In defense of epicycles

Ansgar D. Endress
Department of Psychology, City, University of London, UK

Draft of October 31, 2019

As simpler scientific theories are preferable to more convoluted ones, it is plausible to assume that biological learners are also guided by simplicity considerations when acquiring mental representations, and that formal measures of complexity might indicate which learning problems are harder and which ones are easier. However, the history of science suggests that simpler scientific theories are not necessarily more useful if more convoluted ones make calculations easier. Here, I suggest that a similar conclusion applies to mental representations. Using case studies from perception, associative learning and rule learning, I show that formal measures of complexity critically depend on assumptions about the underlying representational and processing primitives and are generally unrelated to what is actually easy to learn and process in humans. An empirically viable notion of complexity thus need to take into consideration the representational and processing primitives that are available to actual learners even if this leads to formally complex explanations.

Keywords: Language acquisition; perceptual or memory primitives; induction; learning costraints

We like to explain things. Luckily we are not particularly good at it and can thus enjoy stories of gods creating other gods by throwing around godly genitals (Hesiod, 1914) as opposed to less dramatic and "simpler" descriptions of celestial bodies forming other celestial bodies when their dust moves peacefully through space and accretes (Armitage, 2008).

While it might not make for good stories, a particularly prominent metric of the quality of an explanation is its simplicity. This principle is known as Occam's razor: when two theories explain the same phenomenon, the simpler one, requiring fewer assumptions, should be given preference.

As simplicity is a good metric to gauge the quality of *scientific* theories, and as our mental representations are arguably some form of *mental* theory of the world, many authors propose that simplicity might also be a good guiding principle for the kinds of mental representations we entertain and, more generally, for the kinds of inductive biases we have when we learn about the world to establish these mental representations to begin with.

However, the history of science suggests that a formal notion of simplicity is not the only yardstick to judge the quality of a scientific theory. For example, Russo (2000) argues that the often ridiculed Ptolemaic model of planetary motion was an exceptionally efficient computational device:

> Because "epicycles" is still a byword for clumsy and backward attempts at science, we spell out the two reasons why the method was supremely well-adapted to the purposes to which it was put.
>
> First, accounting for the observed motion of planets as the composite of several uniform motions on circular orbits (the first centered on the earth [...] and each of the others, called epicycles, centered on the point obtained on the preceding circumference) is equivalent to a modern expansion in Fourier series, and allows and efficient description of observed data with increasing precision as the number of epicycles grows. [...] Second, since the main computational tool of Hellenistic mathematics was geometric algebra performed with ruler and compass, decomposition into circular motions was the most efficient possible system for computing the observable position of planets. (p. 90/91)

In other words, although heliocentric models have been known for centuries at the time of Claudius Ptolemaeus, the computational tools available to Hellenistic astronomers made a geocentric model efficient for their purposes even though the geocentric model is arguably simpler.

In this paper, I argue that a similar situation arises when it comes to mental representations and the learning thereof. We are endowed with a certain set of computational tools, and any learning problem (as well as the application of what has been learned) needs to make do with the tools we have at our disposal. As a result, a psychologically viable notion of simplicity needs to take into account the computational machinery we have evolved, both for learning and, later on,

for applying what has been learned (i.e., for processing).

Specifically, I will make three claims. First, what is simple from a formal point of view is generally unrelated to what is simple from a learning and processing point of view, presumably because it is easier to learn and process what is more frequent in our environment, which, in turn, makes it more likely that we have evolved machinery to deal with these frequent occurrences. Second, claims that learning is guided by simplicity considerations either misstate mathematical facts or are empirically unsupported. Third, I sketch an empirical concept of simplicity that, with duly overoptimistic expectations about the kinds of progress that might be achievable in our understanding of the mind, might then be used to calculate the computational complexity of operations for human processing.

1 Is formal complexity related to easy of processing?

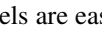
1.1 Extraneous information in inference making

Occam's razor has a venerable history in science. As simpler models are easier to falsify, Popperian, falsification-based scientific paradigms will thrive with simpler hypotheses. But people are not scientists. They indulge in conspiracy theories that require numerous auxiliary hypotheses and like to think that chemicals in the drinking water turn us gay, that 9/11 was an inside job, that Elvis is alive and so forth (though it is a psychologically interesting question whether people will maintain the latter theory indefinitely beyond the expected human lifespan).

People are not only resistant to correction, which might simply be a case of a confirmation bias (e.g., Kahneman, 2011), and doesn't bode too well for refutation-based scientific theories. They also make decisions based on extraneous information. For example, in the conjunction fallacy, people famously consider it more likely for somebody to be a *feminist* bank teller than to be a bank teller *tout court* when told that she is concerned with issues of discrimination and social justice (e.g., Tversky & Kahneman, 1983). This is even though feminist bank tellers are a sub-set of bank tellers in general so that, for any person, the probability of being a bank teller is higher than that of being a feminist bank teller.

Further, people readily fill in gaps in their knowledge when making inferences. This is not a bug of the mind, but rather a feature, as what people are told is generally underspecified. For example, to borrow an example from Sperber and Wilson (1987), a perfectly reasonable answer to the question "Do you want some coffee" is "Coffee will keep me awake." However, this sentence is not an actual answer to the question. Rather, the listener has to infer from extraneous information that is little more than a physiological truism ("coffee will keep me awake") that the actual answer is "No." These kinds of situations are so widespread that entire theories have been devoted to them (e.g., Sperber & Wilson, 1987, 1995), and should abolish any hope that successful communication would be possible if a listener's inferences were "simple" in any relevant sense.

Of course, the tendency of perceivers to have rich interpretations of the world is not limited to high-level phenomena such as decision-making and communication (though they are amply documented in such domains; see e.g. Gigerenzer & Goldstein, 1996; Todd & Gigerenzer, 2000). Rather, they routinely occur in perception as well. In the words of Ramachandran (1991) "perception is essentially a 'bag of tricks' ; [. . .] through millions of years of trial and error the visual system has evolved numerous short-cuts, rules-of-thumb and heuristics which were adopted not for their aesthetic appeal or mathematical elegance but simply because they worked."

For example, if we see two lines of equal length, the simplest inference is probably that they *are* of equal length. But this is not the inference the perceptual system draws: Even though the two lines are of the same length, a line surrounded by arrow tails (e.g., ) is perceived as longer than one surrounded by arrow heads (e.g., ). In this case, however, the reason for the arguably more complex inference is clear: in natural scenes, line segments surrounded by arrow tails tend to be longer than those segments surrounded by arrow heads (Howe & Purves, 2005). Given these environmental statistics, our perceptual system thus evolved to make inferences that are maybe not "simple" in any obvious formal sense (that does not take into consideration historical environmental statistics), but that are adaptive given those environmental statistics.

In the next section, I will argue that similar phenomena generally occur in learning and processing: simplicity in a formal sense is a poor guide towards what is easy to learn or process.

1.2 Simpler isn't simpler

There is no shortage of examples where formally more complex operations are easier for humans. For example, (arithmetic) divisions are hard for humans but easy for computers, while we can do elaborate 3D rotations which require considerable processing power in computers.

Likewise, it is easier to notice that a figure is symmetric (e.g.,) than that it is composed of the copy of two shapes (e.g., ) even though symmetry entails more operations than copying (e.g., Baylis & Driver, 1994, 2001; Bruce & Morgan, 1975; Corballis & Roldan, 1974): In the second figure, the  shape is copied and then translated to the right; in the first figure, the is again copied and then translated, but, crucially, also mirror-reversed. Still, it is easier to detect symmetry than translation.

Such cases suggest that formally "simpler" operations are not necessarily easier to process for humans. But ease of processing is not even constant for a given operation, as the

difficulty to learn or execute a given operation strongly depends on the domain in which it is applied (see Endress, in press, for a review).

This is true even for the simplest operations such as associations. For example, animals readily associate tastes visceral sickness and external events (e.g., sounds and light) with pain. In contrast, it is much more difficult or even impossible to associate taste with pain or external events with sickness (e.g., Garcia & Koelling, 1966; Garcia, Hankins, & Rusiniak, 1974, 1976; see e.g. Domjan, 1983, 2015, for reviews). Evolutionarily speaking, this makes sense of course: Sickness typically results from what we ingest, while physical pain typically has external causes that we can perceive sensorily, which makes the pattern of preferential associations adaptive (see Alberts & Gubernick, 1984; Gemberling & Domjan, 1982; Gemberling, Domjan, & Amsel, 1980; Gubernick & Alberts, 1984, for evidence that it is not learned).

Likewise, if we hear, see or feel two objects frequently occurring together, we form associations between them (e.g., Aslin, Saffran, & Newport, 1998; Conway & Christiansen, 2005; Endress, 2010; Fiser & Aslin, 2002; Saffran, Newport, & Aslin, 1996; Saffran, Johnson, Aslin, & Newport, 1999; Turk-Browne, Jungé, & Scholl, 2005; Turk-Browne & Scholl, 2009). However, this form of associative learning works better for consonants than for vowels (Bonatti, Peña, Nespor, & Mehler, 2005), although the reasons are debated (Bonatti, Peña, Nespor, & Mehler, 2007; Keidel, Jenison, Kluender, & Seidenberg, 2007). Be that as it might, such results suggest that formal complexity is a poor guide to the ease of acquisition and processing of an operation.

A similar conclusion follows from an operation that has helped defining the complexity of phonological processes (e.g., Culy, 1985; Manaster-Ramer, 1986): repetition-patterns. Repetition-patterns are important in many languages. For example, Marshallese uses reduplications for derivative morphology (e.g., "takin" means sock, while "takinkin" means to wear socks; Moravcsik, 1978). In some form or another, reduplication occur in some 85% of the world's languages (Rubino, 2013). Other examples of repetition-patterns used by language include vowel harmony (e.g., Rose & Walker, 2011; Vroomen, Tuomainen, & de Gelder, 1998), the feature repetitions resulting from assimilation rules (e.g., Darcy, Ramus, Christophe, Kinzler, & Dupoux, 2009; Mitterer & Blomert, 2003) and constraints on consonant co-occurrence in Semitic languages such as the Obligatory Contour Principle (e.g., Berent & Shimron, 1997; Frisch, Pierrehumbert, & Broe, 2004; McCarthy, 1986; McCarthy & Prince, 1999).

Despite their importance for language, repetition-patterns can be perceived in many non-linguistic domains and by many non-linguistic animals. For example, even seven-month-old babies notice the repetition in syllable sequences such as *dubaba* generalize it to new items (Marcus, Vijayan, Rao, & Vishton, 1999). When familiarized with syllable sequences such as *ledidi*, *wijeje* and so forth, infants behave as if they were more familiar with novel sequences with novel syllables that share the repetition-pattern (e.g., *bapopo*) than with other novel syllable sequences that do not share the pattern (e.g., *babapo*). Further research revealed that humans and non-human animals can compute repetition patterns not only for speech syllables, but also for tones and visual objects (e.g., Dawson & Gerken, 2009; Endress, Dehaene-Lambertz, & Mehler, 2007; Giurfa, Zhang, Jenett, Menzel, & Srinivasan, 2001; Hauser & Glynn, 2009; Marcus, Fernandes, & Johnson, 2007; Marcus et al., 1999; de la Mora & Toro, 2013; Murphy, Mondragon, & Murphy, 2008; Neiworth, 2013; Pepperberg, 1987; Saffran, Pollak, Seibel, & Shkolnik, 2007; Smirnova, Zorina, Obozova, & Wasserman, 2015; Versace, Spierings, Caffini, Ten Cate, & Vallortigara, 2017; Yamazaki, Suzuki, Inada, Iriki, & Okanoya, 2012, but see Spierings & ten Cate, 2016; van Heijningen, Chen, van Laatum, van der Hulst, & ten Cate, 2013, for evidence that these relations are not equally salient to all species). Further, a sensitivity to repetition-patterns is unlikely to be learned as humans are sensitive to such patterns from birth (e.g., Antell, Caron, & Myers, 1985; Gervain, Berent, & Werker, 2012; Gervain, Macagno, Cogoi, na, & Mehler, 2008), which might be because humans might have a "repetition-detector" that makes such patterns salient (Endress et al., 2007).

However, although the ability to compute such patterns is widely shared across animals and domains, there are some domains where they are easier to learn than in others. For example, adult speakers learn such patterns better when they are carried by vowels than when they carried by consonants, to the extent that they fail to detect the patterns on consonants (e.g., Toro, Bonatti, Nespor, & Mehler, 2008; see Hochmann, Benavides-Varela, Nespor, & Mehler, 2011; Pons & Toro, 2010, for similar results with infants), even when the salience of the vowels is much reduced (Toro, Shukla, Nespor, & Endress, 2008). Further, given that rats learn repetition-patterns just as well on consonants as on vowels (de la Mora & Toro, 2013), the reason for the vowel advantage in humans does not seem to be a gross perceptual difference.

Likewise, adult speakers seem unable to learn repetition-patterns over syntactic categories (Endress & Hauser, 2009). When familiarized with word triplets conforming to either an AAB pattern (noun-noun-verb triplets such as town-leg-choose and verb-verb-noun triplets such as choose-speak-leg) or an ABB pattern (noun-verb-verb or verb-verb-noun), they are unable to decide whether test triplets made from new words were like the familiarization items. They even fail when they are first primed on nouns and verbs, and then told to watch out for a pattern of nouns and verbs (Endress & Hauser, 2009). Again, there does not seem to be any particular formal reason for this failure: participants readily access the categories, learn other sequential regularities

about the categories that do not involve repetitions, and learn repetition-patterns over semantic, non-syntactic categories (e.g., animals and clothes). Endress and Hauser's (2009) conclusion was that repetition-patterns over syntactic categories are simply not in the learner's repertoire because they are not part of any human language, and found some evidence in support of this possibility. Critically, however, participants failed to learn an extremely simple regularity that they should learn based on simplicity considerations, as they can learn it in other context.

Taken together, this evidence thus suggests that the formal simplicity of an operation is a poor guide to how easily it is learned.

2 Are formal notions of simplicity empirically adequate?

In this section, I first consider two related measures of "simplicity" (or rather complexity), one that might be more intuitive for defining the complexity of *operations* to be learned (e.g., grammatical rules) and one more intuitive for the learning of *knowledge structures* (e.g., words). In either case, I will argue that, in the absence of an exhaustive list of the processing and representational primitives used by human learners, neither form of a complexity provides a useful guidance towards the relative ease of a learning problem.

Following this, I will consider a notion of simplicity that focuses on the simplicity of the inferences, and argue that there is no empirical support for it.

2.1 Kolmogorov complexity

Regarding the learning of procedures, one way of defining the "simplicity" of a learning problem is to use Kolmogorov complexity (KC; e.g. Chater, 1996, 1999; Chater & Vitányi, 2003; Pothos & Chater, 2002). KC is basically the length of the shortest program used to describe an object. But of course, the length of a program depends on the language it is written in. According to proponents of the use of KC as a measure of cognitive complexity, this problem is immaterial because, in the words of Chater and Vitányi (2003), "the choice of programming language does not matter [for computing KC], up to a constant additive factor."

However, this statement is misleading. In fact, the *difference* between the program length in different languages is certainly bounded by a constant; that is, if $K_1(x)$ and $K_2(x)$ are the complexities of some procedure in two languages L_1 and L_2, then $|K_1(x) - K_2(x)| < C$, where C is some constant. The reason is simply that, with general-purpose programming languages, one can always write an emulator of L_2 in L_1 (or vice-versa). Once the emulator is written, any piece of code from L_2 will execute in L_1 and the total length of the code is the length of the original L_2 code plus the length of the code representing the emulator when the latter is needed. The constant is thus essentially the code length of the emulator.

Critically, this result does not guarantee that the two objects have the same *relative* KCs in two languages. For example, if L_1 has operations for multiplication and addition but L_2 only for addition, a program using multiplication will be longer than a program using addition in L_2 (because on has to write the multiplication operation first), but both programs will be equally long in L_1. As a result, in the absence of an exhaustive list of the computational primitives available to human learners, it is impossible to formally determine if one learning problem is easier than another one.

Of course, it is possible to turn this approach around and to determine the available primitives from their relative ease of processing (e.g., Katzir, 2015), but this is not how KC is usually employed.

2.2 Minimum Description Length

A closely related measure of complexity is Minimum Description Length (MDL; see Rissanen, 2008 for a basic introduction). For example, when fitting a polynomial through a number of points, a higher degree polynomial will obviously fit the points better, so that we need fewer bits to represent the noise (i.e., the errors from the fit). The description length of the data is thus shorter with higher order polynomials. However, we also need to represent polynomials themselves, and the additional terms of the higher order polynomials make their description length longer. MDL thus seeks a compromise between a short description of the data and a short description of the function describing the data.

This approach has been the basis of one of the first computational model of how infants might learn words from fluent speech (Brent & Cartwright, 1996, though one can also use it to simultaneously a lexicon and rules that are applied to the lexicon, given that output forms are generated from a lexicon using rules; e.g., Katzir, 2015; Rasin & Katzir, 2016). Fluent speech is a continuous signal, according to many authors with few consistent cues to word boundaries (e.g., Aslin et al., 1998; Aslin & Newport, 2012; Conway & Christiansen, 2005; Saffran et al., 1996, 1999, but see e.g. Brentari, González, Seidl, & Wilbur, 2011; Christophe, Gout, Peperkamp, & Morgan, 2003; Endress & Hauser, 2010; Fenlon, Denmark, Campbell, & Woll, 2008). As a result, infants need to figure out where words start and where they end before they can learn the meaning of any word.

Brent and Cartwright (1996) used a MDL-based algorithm similar to standard compression techniques. The basic idea is illustrated in Figure 1, where I show three candidate segmentations of the continuous sequence "Thedogbitesthedog." Each candidate segmentation will take up memory space in three ways. First, we need to reserve memory space for each word in the lexicon. Second, we need to populate this memory space with content, and the longer each word,

the more space it takes up. Third, we need to represent the sequence using the symbols from the lexicon. For our purposes, the optimal segmentation is the one that minimizes the sum of these memory components, though Brent and Cartwright's (1996) model is mathematically more sophisticated.

In the maximal segmentation, each letter is a word in the lexicon. This gives us 9 lexical entries (as there are 9 unique letters in the sequence), each of which takes up a single memory unit. Finally, we need 17 units to represent the 17 letters of the sequence, leading to a total memory score of $9 + 9 + 17 = 35$.

In the minimal segmentation, the entire sequence is stored as a unit in the lexicon. If so, we have a single unit in the lexicon that takes up 17 memory units, while the sequence can be represented with a single symbol, yielding a total memory score of $1 + 17 + 1 = 19$.

Finally, in the intermediate segmentation, we postulate the "words" *bites* and *thedog*. We thus have 2 memory items that take up 11 memory units in total, and that allow us to represent the sequence using only 3 units. The total memory score is thus $2 + 11 + 3 = 16$, and thus the lowest one of the three possible segmentations.

Brent and Cartwright (1996) showed that such an MDL approach successfully recovers many word boundaries. Brent and Cartwright's (1996) critical conclusion was that, given that the algorithm recovered word boundaries, there must have been distributional information that allowed the algorithm to do so and might also allow infant learners to find word boundaries. Critically, they pointed out that their model was not meant to be a psychologically realistic model of infant learning (it is fundamentally a compression algorithm after all) and that their model demonstrated the availability of distributional information that infants might exploit if they have required processing mechanisms.

Here, I add three more general reasons for which MDL-based approaches are unlikely to be good guides to learning. First, as in the complexity case, we don't know what the underlying primitives are. To keep with the example of word segmentations, learners of different languages might well rely different perceptual units (as do their adult counterparts who use stress-based units, syllables or moras, depending on their native language; e.g. Cutler, Mehler, Norris, & Segui, 1986; Cutler & Mehler, 1993; Mehler, Dommergues, Frauenfelder, & Segui, 1981; Otake, Hatano, Cutler, & Mehler, 1993). If the underlying units are different, so we will be the results from a MDL-based learner. For example, when learners encounters their sibling's exclamation "Mama, Papa!" (maybe to express their outrage at the suggestion that they might be simplicity-based learners), it is easy to verify from Brent and Cartwright's (1996) cost function (Equation (2) in Appendix A) that learners using phonemes as basic units would posit the words *Ma* and *Pa*, while learners using syllables as basic units would be undecided between the words *Mama* and *Papa* and the words *Ma* and *Pa*, respectively. Further, even if learners use the same basic units such as syllables or moras, they might not perceive all units equally well depending on their native language (e.g., Dupoux, Pallier, Sebastian, & Mehler, 1997; Polka & Werker, 1994; Werker & Tees, 1984) or perceive extra units that are not present in the speech signal (such as epenthetic vowels; see e.g. Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999).

These problems are more general than the word segmentation problem. For example, if learners in the curve fitting example above have the ability to represent polynomials only up to a degree of 3 (maybe due to some memory limitations), or if they cannot represent polynomials at all, an optimal polynomial of degree 4 would simply not be in their learning repertoire. As a result, without knowing what the underlying representational and processing primitives are, MDL-based approaches do not give us any guidance as to which learning problems might be simpler and which might be harder.

A second problem with MDL-based approaches to learning is that there is no evidence that learners actually optimize the description length. To use the word-segmentation example again, a learner in a word learning experiment might be exposed to a sequence of word repetitions like *dog, dog, dog, . . . , pig, pig, pig, ...* Would such a learner extract the words *dog* and *pig*, or some other units such as *dogdog. . .* and *pigpig. . .* ? In Appendix A, I show that this depends on the specific familiarization: if each word is repeated N times, the optimal units have length $\sqrt{\frac{N}{3 \log_2 5}}$. If each pronunciation of a syllable takes roughly 500 ms, learners should thus extract the units *dogdogdog* and *pigpigpig* after a 1 min familiarization, *dogdogdogdog* and *pigpigpigpig* after a 2 min familiarization, *dogdogdogdogdog* and *pigpigpigpigpig* after a 3 min familiarization, and so forth (see Figure 2), a prediction that seems implausible at best. As a result, MDL-based approaches per se cannot give us any guidance as to the hypotheses learners consider in the absence of independent evidence for the underlying units and for the ability of learners to perform the relevant optimizations (though again, we might use the output of actual learners to decide between different theories of the underlying representations and learning mechanism; Katzir, 2015).

A final problem with MDL-based approaches is that it is not clear that learners really need to optimize the total memory space they use up, given that they have a massive capacity for declarative memory both in the short-term (e.g., Endress & Potter, 2014; Endress & Siddique, 2016) and in the long-term (e.g., Brady, Konkle, Alvarez, & Oliva, 2008; Standing, Conezio, & Haber, 1970; Standing, 1973), with an estimate that a typical college-aged adult knows about 50,000 words (Pinker, 1999). Further, to the extent that grammatical knowledge relies on procedural rather declarative mem-

Segmentation type	Words in lexicon	Sentence representation based on this lexicon	# Words in lexicon	# Letters in lexicon	# Words in sentence	Total
Maximal	b, d, e, g, h, i, o, s, t	9-5-3-2-7-4-1-6-9-3-8-9-5-3-2-7-4	9	9	17	**35**
Minimal	Thedogbitesthedog	1	1	17	1	**19**
Intermediate	bites, thedog	2-1-2	2	11	3	**16**

Figure 1. Illustration of the Minimum Description Length principle for a segmentation of the chunk "The dog bites the dog." The maximal segmentation considers each letter as a word; the minimal segmentation considers the whole input as a single word. Finally, some intermediate segmentations will find recurring units that are larger than single letters and smaller than the entire input. As memory is taken up by representing the items in memory and by representing the sequence, the optimal segmentation that minimizes the total memory storage will be some intermediate segmentation.

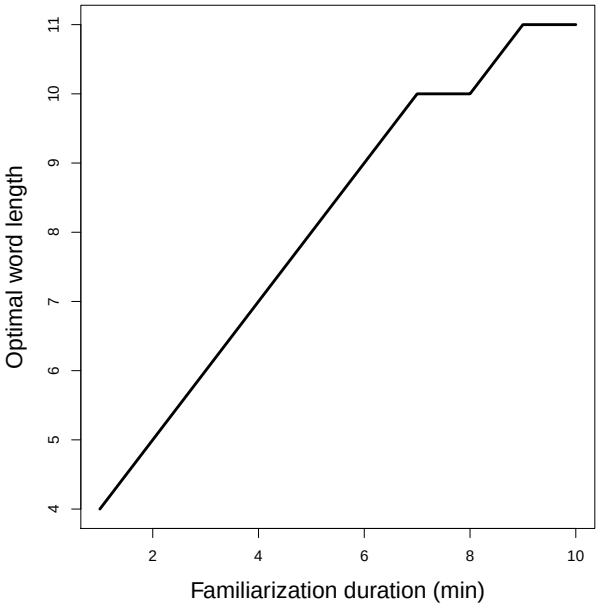


Figure 2. Predicted optimal word length of the extracted units when learners in a word learning experiment are exposed to sequences like *dog, dog, dog, . . . , pig, pig, pig, ...* The prediction assumes that each pronunciation takes about 500 ms.

ory (e.g., Pinker & Ullman, 2002; Ullman et al., 1997; Ullman, 2001), such memory does not appear to be particularly limited either in a species that manages to learn the motor commands associated with speaking, cycling, swimming and playing the Fantasia Contrappuntistica.

2.3 The size principle

An alternative to defining the simplicity of a learning problem is to consider the restrictiveness of the solutions: If we have to choose between two hypotheses that are equally consistent with the examples we have seen, we choose the hypothesis that is more restrictive, a strategy called the *size principle* (Tenenbaum & Griffiths, 2001). A related strategy has been proposed for language acquisition (e.g., Hyams, 1986; Manzini & Wexler, 1987). In some versions of this proposal, humans evolved to acquire language following a sequence of acquisition steps that is consistent with the most restrictive grammar given the input, using specific "triggers" to move from a more restrictive grammar to a more permissive one (e.g., Gibson & Wexler, 1994). The underlying idea is that the triggers allow learners to "conclude" that their current grammars are not general enough, and to adjust them appropriately, while it is unclear how they could even notice that they started out with a grammar that is too general.

In the literature following Tenenbaum and Griffiths (2001), this idea has been applied much more widely in domains ranging from basic probabilistic inference to language acquisition to social cognition. If these demonstrations are convincing, the size principle might provide a simplicity-based metric with which learning problems can be evaluated *in general*. However, I will discuss some of the strongest evidence for the size principle in the domains of word learning, rule learning and probabilistic inference, and argue that these demonstrations provide good examples of what Glymour (2007) called "Ptolemaic Psychology", and that their success relies on numerous auxiliary assumptions. I will further argue that common-sense psychology provides much more straightforward accounts even though it does not provide any general metric for the evaluation of learning problems.

2.3.1 The size principle and word learning (1). Some of the strongest evidence for the size principle comes from Xu and Tenenbaum's (2007b) experiments on word learning (see also Navarro, Dry, & Lee, 2012). They asked how learners assign meaning to novel nouns, and under what condition they would choose a meaning at the subordinate category level (e.g., "Dalmatian"), at the basic-level category level (e.g., "dog"), or at the superordinate category level (e.g., "animal").

Participants were presented with a novel word (e.g., "fep"), and shown one or three examples of the word's meaning (e.g., a Dalmatian). Following this, they were shown

a test screen with potential examples of "feps", and had to select other feps.

The test screen comprised 8 items of each of three superordinate categories (i.e., animals, vegetables and vehicles). Within each category, there were 2 examples of the same subordinate category (e.g., two other Dalmatians), 2 examples of the same basic-level category (e.g., two non-Dalmatian dogs), and 4 examples of the same superordinate category (e.g., four non-dog animals). If participants infer that fep means "animal", they should choose all eight pictures of that category; if they infer that it means "dog," they should choose the four corresponding pictures; and if they conclude that it means "Dalmatian," they should choose the two Dalmatians only.

Results showed that, when familiarized with three Dalmatians, participants concluded that fep meant "Dalmatian." When familiarized with one Dalmatian and two other dogs, they concluded that fep meant "dog;" and when familiarized with one Dalmatian and two non-dog animals, they concluded that fep meant "animal".

These results are certainly consistent with the size principle. After all, there are more animals in the world than there are dogs, and there are more dogs than there are Dalmatians. Hence, if learners opt for the most restrictive inference, they should opt for a subordinate meaning when this it is consistent with the data as this is the most restrictive one, and chose basic level or superordinate levels only when required by the data.

However, it is almost certain that learners do not apply the size principle, simply because it does not scale up to natural language acquisition. In fact, natural language learners are rarely shown test screens explicitly providing them with the 9 possible meanings of a novel word as well as with the number of elements of each category, raising the question of how learners could possibly estimate the number of elements of a category — and even if they had access to this information, whether they could process it. In fact, there are an estimated 70 million dogs in the United States (American Veterinary Medical Society 2012 U.S. Pet Ownership & Demographics Sourcebook, retrieved on 8/16/2019 from `https://www.avma.org/KB/Resources/Statistics/Pages/Market-research-statistics-US-pet-ownership.aspx`). It is an entirely open question whether infants can process numbers of this magnitude, or which other information they might possibly exploit according to Xu and Tenenbaum (2007b).

In contrast, these results follow directly from standard approaches to word learning (e.g., Medina, Snedeker, Trueswell, & Gleitman, 2011; Stevens, Gleitman, Trueswell, & Yang, 2017): Learners acquire a word whenever they encounter a situation that is conducive for inferring its meaning. If subsequent occurrences are consistent with this guess, they stick with it, and revise it if they are not. By default, learners might assume that word meanings correspond to a basic-level category (e.g., Markman & Hutchinson, 1984; Waxman & Markow, 1995). If they see an example that is not part of the basic level category (e.g., a cat, which is an animal but not a dog), they might revise their guess and opt for a superordinate category interpretation, similar to how the "triggers" discussed above (e.g., Gibson & Wexler, 1994) might lead learners to move to more general hypotheses. Conversely, if they consistently see "bad" examples of a basic-level category (e.g., because Dalmatians are presumably not particularly prototypical dogs; Emberson, Misyak, Schwade, Christiansen, & Goldstein, 2019), or if the variability of the exemplars is less than expected from a basic-level category (e.g., because learners are not shown the best examples), learners might opt for a subordinate interpretation. This model thus accounts for Xu and Tenenbaum's (2007b) data, without making any use of the size principle.[1]

That being said, Xu and Tenenbaum (2007b) report another result that, at first sight, seems to provide strong evidence for the size principle. Specifically, they show that, when participants are familiarized with a single Dalmatian, they conclude that fep means "dog;" in contrast, when familiarized with three Dalmatians, they conclude that fep means "Dalmatian." However, as mentioned above, participants might have a tendency to use basic level categories, and conclude that the lone Dalmatian is an exemplar of the "dog." When shown three Dalmatians, however, they might be surprised that all of the dogs are Dalmatian (e.g., because the examples are particularly prototypical dogs), and change their inference accordingly.[2]

[1] This explanation predicts that learners should be less willing to entertain a subordinate interpretation when the examples are more prototypical for the basic level category even if they all come from the same subordinate category. For example, if they are shown German shepherds rather than Dalmatians, a subordinate interpretation might be less available.

[2] In contrast, according to Xu and Tenenbaum's (2007b) formal explanation, the likelihood of each category given all exemplars is the product of the likelihoods of each category given the individual exemplars. Importantly, one of the factors in the individual likelihoods comes from the size principle, and is inversely proportional to the number of items in the category. As a result, with more exemplars, the influence of the set size is more pronounced, and should favor smaller categories (e.g., subordinate categories if these are consistent with the data). However, this is true for any probability smaller than 1 that is raised to a power corresponding to the number of exemplars. To make this point, I show in Appendix B that Xu and Tenenbaum's (2007b) results can also be explained based on the similarity between the training items and the test items. If the similarity score is converted to a probability score, the same qualitative predictions follow. That is, the a priori likelihood of a meaning is largest when the similarity between the training examples and the other items to which the meaning applies is largest; as such, this probability score favors small (i.e., subordinate) categories. Hence,

2.3.2 The size principle and word learning (2). Like Xu and Tenenbaum (2007b), Xu and Tenenbaum (2007a) asked how learners choose between subordinate category and basic-level meanings. Participants (adults and 3-to 4-year-olds) were presented with a display showing items from two categories. The categories were defined by their shapes and were spatially grouped together. Each category comprised three subordinate categories of 4 items each. Items in different subordinate categories shared their shape, but differed in texture.

Learning occurred in one of two conditions. In the teacher-driven condition, the experimenter pointed to three items from the same subordinate category and labeled them with a novel word (e.g., a blicket). In the learner-driven condition, the experimenter labeled only one object, and then encouraged participants to point to two *other* blickets. Results showed that all but one participant pointed to two other items from the same subordinate category.

Following this, participants were pointed to 5 more objects and asked whether these were blickets as well. Results showed that participants in the teacher-driven condition were more likely to infer that "blicket" refered to the subordinate category than participants in the learner-driven condition. Xu and Tenenbaum's (2007a) explain the preference for the subordinate category in the teacher-driven condition as follows. Since the teacher knows the meaning of the word, she will choose example objects to which this meaning applies. Hence, due to the size principle, if all examples are consistent with a subordinate meaning, a subordinate meaning should be preferred, because the number of items in that category is lower than that in a basic-level category. In the learner-driven condition, in contrast, the learner does not know the meaning of the word; hence, the size principle does not apply to favor smaller categories, such that the learner should be less likely to infer a subordinate-level meaning than in the teacher-driven condition.

At first sight, Xu and Tenenbaum's (2007a) assumption that the size principle applies only in the teacher-driven condition seems to contradict Xu and Tenenbaum's (2007b) assumption that the size-principle does apply in the absence of a teacher. Maybe more critically, if it is true that, for the size principle to apply, one needs an experimenter (or caretaker) who patiently points to all instances in the extension of a label, it would seem questionable whether the size principle has any relevance for language acquisition at all. Further, the size principle would not apply to the acquisition of verbs — for which deictic reference is much less felicitous and much more ambiguous than for nouns (e.g., "Look at this, that's digesting!")

Leaving aside this caveat, Xu and Tenenbaum's (2007a) results are actually inconsistent with their conclusions. In fact, all but one participant in the learner-driven condition selected items from the same subordinate category when asked to find other blickets; given that Xu and Tenenbaum's (2007a) model predicts that participants in the learner-driven condition should favor a basic-level interpretation, and that there are more than twice as many candidate blickets from different subordinate categories, one would expect them to preferentially choose items from different subordinate categories.

However, there is a simple alternative interpretation. Xu and Tenenbaum's (2007a) claims notwithstanding, participants clearly have a tendency to choose a subordinate-level interpretation, maybe because they are presented with novel non-sense objects that might not be readily assigned to conceptual categories (Callanan, Repp, McCarthy, & Latzke, 1994). In the teacher-driven condition, participants might just stick with this interpretation. In the learner-driven condition, in contrast, they might opt for the basic-level interpretation for purely pragmatic reasons after they initially chose a subordinate interpretation. As mentioned above, participants were asked to decide which *other* objects were blickets only after they had (correctly) identified two further blickets. Plausibly, the teacher pointing to further objects, asking whether they were blickets as well, might have given participants the impression that their initial interpretation (at the subordinate level) was not general enough, and that the experimenter expected a (more general) basic-level interpretation. If so, the difference between the teacher-driven condition and the learner-driven condition might be due to pragmatic factors and unrelated to the size principle.

2.3.3 The size principle and rule-learning. Frank and Tenenbaum (2011) use the size principle to explain how infants might learn the repetition-patterns discussed above. They propose that, when infants have to choose between multiple regularities that are consistent with examples they have heard, they choose the one that has fewer potential items conforming to it.

According to their model, infants might encounter a total of three syllables. Before encountering any syllable triplet, infants know that the three syllables allow for a total 27 triplets, that 6 of these triplets follow an ABB pattern (e.g., *pu-li-li*), that 3 of these triplets follow an AAA pattern (where all three syllables are identical), as well as the number of triplets that would conform to any conceivable rule. They

---

when it is raised to the third power (due to the three examples), the preference for the subordinate category will be more pronounced. As a result, what guarantees the narrowing of the generalizations between one and three exemplars is not the size principle but rather the rest of Xu and Tenenbaum's (2007b) formalism.

A similar conclusion applies to the results reported by Navarro et al. (2012). "Narrowing" of the inferences occurs simply due to raising probabilities to a power corresponding to the number of examples, but not due to the size principle per se. As such, neither Xu and Tenenbaum's (2007b) nor Navarro et al.'s (2012) results provide evidence for computations involving the size principle.

then use the number of triplets that are consistent with each rule to choose among possible generalizations.

Frank and Tenenbaum (2011) applied this size-principle-based model to a variety of infant rule-learning experiments. However, in addition to the *prima facie* implausibility of the model, Endress (2013) showed that the models made incorrect predictions (e.g., that a change from human syllables to monkey vocalizations should be *less* salient than a relatively subtle change from *AAB* patterns to *ABB* patterns), assumed that infants can process about 900 triplets per second, made predictions that were subsequently refuted (Gervain & Endress, 2017), assumed that infants have severe perceptual problems in some phases of an experiment and perfect perception in other phases, used model parameters that led their model to contradict the experimental data when the parameters were used in psychologically meaningful ways or wired in the phenomenon they sought to explain (Endress, 2013; see Frank (2013) and Endress (2014) for discussion). It thus seems that an account based on the size principle is unlikely to explain infant rule learning.

2.3.4 The size principle and probabilistic inference. Gweon, Tenenbaum, and Schulz (2010) presented 15-months-olds with a transparent box containing blue and yellow balls (see also Denison, Reed, & Xu, 2013). The experimenter then removed a variable number of *blue* balls from the box and demonstrated that they squeaked upon squeezing them. Following this, infants were handed a *yellow* ball. Gweon et al. (2010) asked how likely infants were to conclude that this ball squeaked as well. The dependent measure was whether, and how often, infants would squeeze the yellow ball.

The critical manipulations were (1) how many balls the experimenter picked from the box (1 or 3) and (2) whether the majority of the balls in the box was blue or yellow: In blue majority populations, 75% of the balls were blue, while, in yellow majority populations, only 25% of the balls were blue.

When the experimenter extracted *three* blue balls from a blue majority population, infants squeezed the yellow ball more often than when the blue balls came from a yellow majority population. In contrast, when only a *single* blue ball was extracted from a yellow majority population, infants squeezed the yellow ball as much as when three blue balls were extracted from a blue majority population. In a crucial control condition, three blue balls were ostensibly drawn by chance from yellow majority population. In that condition, infants did not suppress squeezing the yellow ball.

To explain their data, Gweon et al. (2010) propose that infants consider the four possibilities spanned by two factors: (i) Is the teacher is cooperative and picks the balls only from the squeaky ones, or is she nasty, and picks from all balls irrespectively of squeakiness? (ii) Are all balls squeaky, or only the blue ones? Infants then compute likelihoods of the results of the experimenter's actions according to all four possible scenarios, and compare these likelihoods to decide whether or not to squeeze the yellow ball. This likelihood ratio is given by, with β being the proportion of blue balls and α being a parameter that is irrelevant for the current purposes:

$$L_\alpha = \frac{\beta^n}{\alpha + (1-\alpha)\beta^n}, \quad \alpha \in [0,1], \beta \in]0,1]. \tag{1}$$

Irrespective of whether it is plausible that infants have the processing abilities to deal with such a complex model, it is actually inconsistent with the data, for two reasons. First, the model always concludes that it is more likely that only blue balls are squeaky. In fact, it is easy to see that $L_\alpha = 1$ for $\alpha = 0$ or $\beta = 1$, and that $L_\alpha < 1$ for all $\alpha > 0$ and $\beta < 1$. Hence, Gweon et al.'s (2010) model predicts that infants should *never* squeeze the yellow ball at all.[3]

Second, the model assumes that it's more fun to squeeze squeaky blue balls than to squeeze squeaky yellow balls, and that infants thus desire to find *blue* squeaky balls. A plausible alternative hypothesis is that infants are interested in squeaky balls irrespective of their color. As shown in Appendix C and in Figure 3, a version of Gweon et al.'s (2010) model that assumes that infants care about squeaky balls *irrespective* of their color predicts that infants should be more likely to squeeze the yellow ball in *yellow* majority populations, which is just the situation where infants are *less* likely to squeeze it. (However, in contrast to Gweon et al.'s (2010) model, the new model accounts for the fact infants squeeze the yellow ball in the first place.) The "success" of Gweon et al.'s (2010) model is not due to the size principle, but rather to extraneous assumptions about what infants are most interested in: They need to assume that squeaky blue balls are more fun than squeaky yellow balls.

In contrast to this model, there is a much simpler explanation. By default, infants might tend to squeeze the balls, because shape is a better predictor of function (i.e., squeaking) than color (e.g., Bloom, 1996; Brown, 1990; Hauser, 1997), and because squeezing it does not entail a huge cost. However, infants can also detect non-random behavior of an agent; they know that drawing three blue balls out of a box of mostly yellow balls is unlikely (Téglás, Girotto, Gonzalez, & Bonatti, 2007) and can use this ability to detect non-random behavior in agents (e.g., Kushnir, Xu, & Wellman, 2010; though they sometimes expect random behavior from agents as well; Tauzin & Gergely, 2019). Further, infants know that humans are often communicative and might even courageously attempt to "teach them" (e.g., Csibra & Gergely,

[3] It is easy to see that, irrespective of the proportion of blue balls, this effect should be more pronounced when more balls are drawn from the container, and that, eventually L_α goes to 0. Indeed, the partial derivative $\partial_n L_\alpha = \frac{\alpha\beta^n \ln\beta}{(\alpha+(1-\alpha)\beta^n)^2}$ is strictly smaller than 0 since $\ln\beta$ is smaller than 0 for $\beta < 1$. Further, $\lim_{n\to\infty} L_\alpha = 0$.

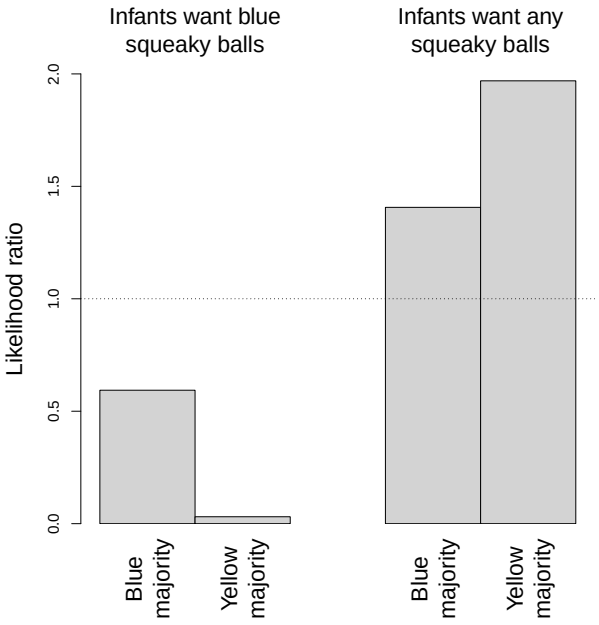


Figure 3. Predictions of Gweon et al.'s (2010) original model assuming that infants like to find squeaky *blue* balls (left) and a modified model where infants are interested in squeaky balls irrespective of color (right) when 3 balls are extracted from the container. In Gweon et al.'s (2010) model, infants are more likely to squeeze a yellow ball in when the container holds a majority of blue balls; in the modified model, infants are more likely to squeeze a yellow ball when the container holds a majority of yellow balls. See Appendix C for details.

2009). Hence, they might detect the non-random behavior of the agent, assume that the agent has a reason to behave in non-random ways, and imitate her more closely only in the condition where the agent shows clear non-random behavior. This idea accounts for all of Gweon et al.'s (2010) data. A similar account applies to Denison et al.'s (2013) data.

Taken together, these data thus do not provide support for another simplicity-based guide to learning: the size principle. Rather, they have alternative explanations based on simple psychological considerations, raising the question of what might constitute an empirically adequate guide to learning.

3 Towards an empirical notion of simplicity

In this paper, I first reviewed data suggesting that people do not necessarily draw the simplest inferences and that, generally speaking, the formal complexity of an operation is not necessarily a good measure of what is easy to process or learn for actual humans. I then showed that prominent formal metrics of simplicity are unlikely to help in this respect. Claims to the contrary notwithstanding, approaches that rely on minimizing the code or description length of a problem crucially depend on assumptions about the underlying representational and processing primitives, and as the empirical examples above show, even when we have a sense about what the underlying computational primitives might be, certain operations are easier to perform in some domains than in others, again for no obvious formal reason. Finally, approaches such as the size principle that quantify the simplicity of the consequences of an inference have, to the best of my knowledge no empirical support, and strongly depend on what the underlying representations are assumed to be.

Together, these examples thus suggest that, for any empirically viable notion of "simplicity", we need to determine empirically what the underlying representational, learning and processing primitives are before any notion of simplicity can be fruitfully deployed. Once (and if) such a list of primitives becomes available, it will become possible to calculate the complexity of a learning or processing problem for an actual biological learning.

From an evolutionary point of view, this conclusion is utterly unsurprising. Just as it is impossible to know that a geocentric model of the solar system is computationally efficient (even when "simpler" heliocentric models are available) without knowing the processing constraints of Hellenistic astronomers (Russo, 2000), it is impossible to know what is simple for humans to infer and process without knowing the (evolutionary) history of their representational and processing abilities. Inductive biases might not follow simplicity prescriptions when considered by themselves, but they might be particularly conducive for learning the regularities that need to be learned given the (linguistic) environment in which they are learned (i.e., the input), just as we preferentially learn ecologically relevant associations. In some cases, models with epicycles are thus the better models.

4 References

Alberts, J. R., & Gubernick, D. J. (1984). Early learning as ontogenetic adaptation for ingestion by rats. *Learn Motiv*, *15*(4), 334 - 359. doi: 10.1016/0023-9690(84)90002-X

Antell, S., Caron, A., & Myers, R. (1985). Perception of relational invariants by newborns. *Developmental Psychology*, *21*(6), 942–948.

Armitage, P. (2008). Planetary formation and migration. *Scholarpedia*, *3*(3), 4479. (revision #186389) doi: 10.4249/scholarpedia.4479

Aslin, R. N., & Newport, E. L. (2012). Statistical learning. *Current Directions in Psychological Science*, *21*(3), 170-176. doi: 10.1177/0963721412436806

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Baylis, G. C., & Driver, J. (1994). Parallel computation of symmetry but not repetition in single visual objects. *Visual Cognition*, *1*, 337–400.

Baylis, G. C., & Driver, J. (2001). Perception of symmetry and repetition within and across visual shapes: Part-descriptions and object-based attention. *Visual Cognition*, *8*(2), 163–196.

Berent, I., & Shimron, J. (1997). The representation of Hebrew words: Evidence from the obligatory contour principle. *Cognition*, *64*(1), 39-72.

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, *60*(1), 1–29.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*(8), 451-459.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2007). On consonants, vowels, chickens, and eggs. *Psychological Science*, *18*(10), 924–925. doi: 10.1111/j.1467-9280.2007.02002.x

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14325–14329. doi: 10.1073/pnas.0803390105

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1-2), 93-125.

Brentari, D., González, C., Seidl, A., & Wilbur, R. (2011). Sensitivity to visual prosodic cues in signers and nonsigners. *Language and Speech*, *54*(1), 49–72.

Brown, A. L. (1990). Domain-specific principles affect learning and transfer in children. *Cognitive Science*, *14*(1), 107–133.

Bruce, V. G., & Morgan, M. J. (1975). Violations of symmetry and repetition in visual patterns. *Perception*, *4*(3), 239–249.

Callanan, M. A., Repp, A. M., McCarthy, M. G., & Latzke, M. A. (1994). Children's hypotheses about word meanings: is there a basic level constraint? *Journal of Experimental Child Psychology*, *57*(1), 108–138. doi: 10.1006/jecp.1994.1006

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*(3), 566-81.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *52*(2), 273–302.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19-22.

Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. (2003). Discovering words in the continuous speech stream: the role of prosody. *Journal of Phonetics*, *31*(3-4), 585–598. doi: 10.1016/S0095-4470(03)00040-8

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(1), 24–39. doi: 10.1037/0278-7393.31.1.24

Corballis, M. C., & Roldan, C. E. (1974). On the perception of symmetrical and repeated patterns. *Perception and Psychophysics*, *16*(1), 136–142.

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

Culy, C. (1985). The complexity of the vocabulary of bambara. *Linguistics and Philosophy*, *8*(3), 345–351. doi: 10.1007/BF00630918

Cutler, A., & Mehler, J. (1993). Mora or syllable? speech segmentation in japanese. *Journal of Memory and Language*, *32*(2), 258–278.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of french and english. *Journal of Memory and Language*, *25*(4), 385–400.

Darcy, I., Ramus, F., Christophe, A., Kinzler, K., & Dupoux, E. (2009). Phonological knowledge in compensation for native and non-native assimilation. In F. Kügler, C. Féry, & R. van de Vijver (Eds.), *Variation and gradience in phonetics and phonology* (pp. 265–309). Berlin: Mouton De Gruyter.

Dawson, C., & Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition*, *111*(3), 378–382. doi: 10.1016/j.cognition.2009.02.010

de la Mora, D. M., & Toro, J. M. (2013). Rule learning over consonants and vowels in a non-human animal. *Cognition*, *126*(2), 307–312. doi: 10.1016/j.cognition.2012.09.015

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: evidence from 4.5- and 6-month-olds. *Developmental Psychology*, *49*(2), 243–249. doi: 10.1037/a0028278

Domjan, M. (1983). Biological constraints on instrumental and classical conditioning: Implications for general process theory. *Psychology of Learning and Motivation*, *17*, 215 - 277. doi: 10.1016/S0079-7421(08)60100-0

Domjan, M. (2015). The Garcia-Koelling selective association effect: A historical and personal perspective. *International Journal of Comparative Psychology*, *28*.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1568-1578.

Dupoux, E., Pallier, C., Sebastian, N., & Mehler, J. (1997). A destressing "deafness" in french? *Journal of Memory and Language*, *36*(3), 406 - 421. doi: 10.1006/jmla.1996.2500

Emberson, L. L., Misyak, J. B., Schwade, J. A., Christiansen, M. H., & Goldstein, M. H. (2019). Comparing statistical learning across perceptual modalities in infancy: An investigation of underlying learning mechanism(s). *Developmental Science*, *22*(6), e12847. doi: 10.1111/desc.12847

Endress, A. D. (2010). Learning melodies from non-adjacent tones. *Acta Psychologica*, *135*(2), 182–190.

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. *Cognition*, *127*(2), 159–176. doi: 10.1016/j.cognition.2012.11.014

Endress, A. D. (2014). How are Bayesian models really used? *Cognition*, *130*(1), 81–84. doi: 10.1016/j.cognition.2013.09.003

Endress, A. D. (in press). Duplications and domain-generality. *Psychological Bulletin*.

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*(3), 577–614.

Endress, A. D., & Hauser, M. D. (2009). Syntax-induced pattern deafness. *Proceedings of the National Academy of Sci-*

ences of the United States of America, *106*(49), 21001-6. doi: 10.1073/pnas.0908963106

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177-199.

Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology. General*, *143*(2), 548–65. doi: 10.1037/a0033934

Endress, A. D., & Siddique, A. (2016). The cost of proactive interference is constant across presentation conditions. *Acta Psychologica*, *170*, 186 - 194. doi: dx.doi.org/10.1016/j.actpsy.2016.08.001

Fenlon, J., Denmark, T., Campbell, R., & Woll, B. (2008). Seeing sentence boundaries. *Sign Language & Linguistics*, *10*(2), 177-200.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(24), 15822-6. doi: 10.1073/pnas.232472899

Frank, M. C. (2013). Throwing out the bayesian baby with the optimal bathwater: Response to. *Cognition*, *128*(3), 417–423. doi: 10.1016/j.cognition.2013.04.010

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371. doi: 10.1016/j.cognition.2010.10.005

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language Linguistic Theory*, *22*(1), 179–228.

Garcia, J., Hankins, W. G., & Rusiniak, K. W. (1974). Behavioral regulation of the milieu interne in man and rat. *Science*, *185*(4154), 824-31.

Garcia, J., Hankins, W. G., & Rusiniak, K. W. (1976). Flavor aversion studies. *Science*, *192*, 265–267.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science*, *4*(1), 123–124. doi: 10.3758/BF03342209

Gemberling, G. A., & Domjan, M. (1982). Selective associations in one-day-old rats: taste-toxicosis and texture-shock aversion learning. *Journal of Comparative and Physiological Psychology*, *96*, 105–113.

Gemberling, G. A., Domjan, M., & Amsel, A. (1980). Aversion learning in 5-day-old rats: taste-toxicosis and texture-shock associations. *Journal of Comparative and Physiological Psychology*, *94*, 734–745.

Gervain, J., Berent, I., & Werker, J. F. (2012). Binding at birth: The newborn brain detects identity relations and sequential position in speech. *Journal of Cognitive Neuroscience*, *24*(3), 1–11.

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

Gervain, J., Macagno, F., Cogoi, S., na, M. P., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(37), 14222-7. doi: 10.1073/pnas.0806530105

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*(3), 407–454.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650-69.

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., & Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature*, *410*(6831), 930-3. doi: 10.1038/35073582

Glymour, C. (2007). Bayesian Ptolemaic psychology. In W. Harper & G. Wheeler (Eds.), *Probability and inference: Essays in honour of Henry E. Kyburg, Jr.* London, UK: College Publications.

Gubernick, D. J., & Alberts, J. R. (1984). A specialization of taste aversion learning during suckling and its weaning-associated transformation. *Dev Psychobiol*, *17*, 613–628. doi: 10.1002/dev.420170605

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(20), 9066–9071. doi: 10.1073/pnas.1003095107

Hauser, M. D. (1997). Artifactual kinds and functional design features: what a primate understands without language. *Cognition*, *64*(3), 285–308.

Hauser, M. D., & Glynn, D. (2009). Can free-ranging rhesus monkeys (Macaca mulatta) extract artificially created rules comprised of natural vocalizations? *Journal of Comparative Psychology*, *123*(2), 161-7. doi: 10.1037/a0015584

Hesiod. (1914). *The Homeric Hymns and Homerica* (Vol. 57). Cambridge, MA: Harvard University Press.

Hochmann, J.-R., Benavides-Varela, S., Nespor, M., & Mehler, J. (2011). Consonants and vowels: different roles in early language acquisition. *Developmental Science*, *14*(6), 1445–1458. doi: 10.1111/j.1467-7687.2011.01089.x

Howe, C. Q., & Purves, D. (2005). The Müller-Lyer illusion explained by the statistics of image-source relationships. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1234-9. doi: 10.1073/pnas.0409314102

Hyams, N. (1986). *Language acquisition and the theory of parameters*. Dordrecht: D. Reidel.

Kahneman, D. (2011). *Thinking, fast and slow*. London, UK: Penguin books.

Katzir, R. (2015). A cognitively plausible model for grammar induction. *Journal of Language Modelling*, *2*(2), 213. doi: 10.15398/jlm.v2i2.85

Keidel, J. L., Jenison, R. L., Kluender, K. R., & Seidenberg, M. S. (2007). Does grammar constrain statistical learning? Commentary on Bonatti, Peña, Nespor, and Mehler (2005). *Psychological Science*, *18*(10), 922-3. doi: 10.1111/j.1467-9280.2007.02001.x

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134–1140. doi: 10.1177/0956797610376652

Manaster-Ramer, A. (1986). Copying in natural languages, context-freeness, and queue grammars. In *Proceedings of the 24th annual meeting on association for computational linguistics* (pp. 85–89). Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.3115/981131.981145

Manzini, M. R., & Wexler, K. (1987). Parameters, binding theory, and learnability. *Linguistic Inquiry*, *18*(3), pp. 413-444.

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological Science*, *18*(5), 387-91. doi: 10.1111/j.1467-9280.2007.01910.x

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77-80.

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, *16*(1), 1 - 27. doi: 10.1016/0010-0285(84)90002-1

McCarthy, J. J. (1986). Ocp effects: Gemination and antigemination. *Linguistic Inquiry*, *17*(2), 207–263.

McCarthy, J. J., & Prince, A. (1999). Faithfulness and identity in prosodic morphology. In R. Kager, H. van der Hulst, & W. Zonneveld (Eds.), *The prosody morphology interface* (pp. 218–309). Cambridge: Cambridge University Press.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(22), 9014–9019. doi: 10.1073/pnas.1105040108

Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, *20*(3), 298–305.

Mitterer, H., & Blomert, L. (2003). Coping with phonological assimilation in speech perception: evidence for early compensation. *Perception & psychophysics*, *65*, 956–969.

Moravcsik, E. (1978). Reduplicative constructions. In J. H. Greenberg (Ed.), *Universals of human language: Word structure* (Vol. 3, pp. 297–334). Stanford, CA: Stanford University Press.

Murphy, R. A., Mondragon, E., & Murphy, V. A. (2008). Rule Learning by Rats. *Science*, *319*(5871), 1849-1851. doi: 10.1126/science.1151564

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223. doi: 10.1111/j.1551-6709.2011.01212.x

Neiworth, J. J. (2013). Chasing sounds. *Behavioural Processes*, *93*, 111–115. doi: 10.1016/j.beproc.2012.11.009

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, *32*, 258–258.

Pepperberg, I. M. (1987). Acquisition of the same/different concept by an african grey parrot (psittacus erithacus): Learning with respect to categories of color, shape, and material. *Animal Learning & Behavior*, *15*(4), 423–432. doi: 10.3758/BF03205051

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456-463.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology. Human Perception and Performance*, *20*, 421–435.

Pons, F., & Toro, J. M. (2010). Structural generalizations over consonants and vowels in 11-month-old infants. *Cognition*, *116*(3), 361–367. doi: 10.1016/j.cognition.2010.05.013

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26*(3), 303-343. doi: 10.1016/S0364-0213(02)00064-2

Ramachandran, V. (1991). Interactions between motion, depth, color and form: The utilitarian theory of perception. In C. Blakemore, K. Adler, & M. Pointoin (Eds.), *Vision: Coding and efficiency* (pp. 346–360). New York: Cambridge University Press. doi: doi:10.1017/CBO9780511626197.033

Rasin, E., & Katzir, R. (2016). On evaluation metrics in optimality theory. *Linguistic Inquiry*, *47*(2), 235-282. doi: 10.1162/LING_a_00210

Rissanen, J. (2008). Minimum description length. *Scholarpedia*, *3*(8), 6727. (revision #91501) doi: 10.4249/scholarpedia.6727

Rose, S., & Walker, R. (2011). Harmony systems. In J. Goldsmith, J. Riggle, & A. C. Yu (Eds.), *The handbook of phonological theory* (2nd ed., p. 240-290). Oxford, UK: John Wiley & Sons. doi: 10.1002/9781444343069.ch8

Rubino, C. (2013). Reduplication. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Russo, L. (2000). *The forgotten revolution. How science was born in 300 BC and why it had to be reborn.* Berlin, Germany: Springer.

Saffran, J. R., Johnson, E., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27-52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–21.

Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: infant rule learning is not specific to language. *Cognition*, *105*(3), 669-80. doi: 10.1016/j.cognition.2006.11.004

Smirnova, A., Zorina, Z., Obozova, T., & Wasserman, E. (2015). Crows spontaneously exhibit analogical reasoning. *Current Biology*, *25*, 256–260. doi: 10.1016/j.cub.2014.11.063

Sperber, D., & Wilson, D. (1987). Précis of relevance: Communication and Cognition. *Behavioral and Brain Sciences*, *10*(4), 697-710.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford: Blackwell.

Spierings, M. J., & ten Cate, C. (2016). Budgerigars and zebra finches differ in how they generalize in an artificial grammar learning experiment. *Proceedings of the National Academy of Sciences*, *113*(27), E3977-E3984. doi: 10.1073/pnas.1600483113

Standing, L. (1973). Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology*, *25*(2), 207–222.

Standing, L., Conezio, J., & Haber, R. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science*, *19*(2), 73-74.

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive Science*, *41*(S4), 638-676. doi: 10.1111/cogs.12416

Tauzin, T., & Gergely, G. (2019). Variability of signal sequences in turn-taking exchanges induces agency attribution in 10.5-moolds. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 15441–15446. doi: 10.1073/pnas.1816709116

Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(48), 19156–19159. doi: 10.1073/pnas.0700271104

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–40; discussion 652-791.

Todd, P., & Gigerenzer, G. (2000). Précis of Simple heuristics that make us smart. *Behavioral and Brain Sciences*, *23*(5), 727-41; discussion 742-80.

Toro, J. M., Bonatti, L., Nespor, M., & Mehler, J. (2008). Finding words and rules in a speech stream: functional differences between vowels and consonants. *Psychological Science*, *19*, 137–144.

Toro, J. M., Shukla, M., Nespor, M., & Endress, A. D. (2008). The quest for generalizations over consonants: asymmetries between consonants and vowels are not the by-product of acoustic differences. *Perception and Psychophysics*, *70*(8), 1515–1525. doi: 10.3758/PP.70.8.1515

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology. General*, *134*(4), 552-64. doi: 10.1037/0096-3445.134.4.552

Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(1), 195–202.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 293–315.

Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, *2*(10), 717-26. doi: 10.1038/35094573

Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., Growdon, J., Koroshetz, W., & Pinker, S. (1997). A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience*, *9*, 266–76.

van Heijningen, C. A. A., Chen, J., van Laatum, I., van der Hulst, B., & ten Cate, C. (2013). Rule learning by zebra finches in an artificial grammar learning task: which rule? *Animal Cognition*, *16*(2), 165–175. doi: 10.1007/s10071-012-0559-x

Versace, E., Spierings, M. J., Caffini, M., Ten Cate, C., & Vallortigara, G. (2017). Spontaneous generalization of abstract multimodal patterns in young domestic chicks. *Animal cognition*, *20*, 521–529. doi: 10.1007/s10071-017-1079-5

Vroomen, J., Tuomainen, J., & de Gelder, B. (1998). The roles of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language*, *38*(2), 133–149.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302. doi: 10.1006/cogp.1995.1016

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

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297. doi: 10.1111/j.1467-7687.2007.00590.x

Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245 - 272. doi: 10.1037/0033-295X.114.2.245

Yamazaki, Y., Suzuki, K., Inada, M., Iriki, A., & Okanoya, K. (2012). Sequential learning and rule abstraction in Bengalese finches. *Animal Cognition*, *15*, 369–377. doi: 10.1007/s10071-011-0462-x

Appendix A

Segmentation of repeated words in Brent and Cartwright's (1996) model

For each candidate segmentation, Brent and Cartwright's (1996) model calculates a cost according the following cost function:

$$\begin{aligned} f(S) = \quad & 3\,|\text{TYPES(S)}| + \\ & \log_2 P\left(\sum_{w\in\text{TYPES(S)}} l(w)\right) + \\ & |\text{TOKENS(S)}| \times H(S) \end{aligned} \tag{2}$$

where S is a segmentation, P is the number of phonemes in the input alphabet, $l(w)$ is the length of word w, TYPES(S) is the set of (unique) words in the segmentation S, |TOKENS(S)| is the total number of words in segmentation S, |.| is the number of elements in the set . and $H(S)$ is the entropy of the (relative) word frequencies

$$H(S) = -\sum_{w\in\text{TYPES(S)}} p(w)\log_2 p(w). \tag{3}$$

The first term of the cost function (2) prevents an excessive number of words in the lexicon while the second term penalizes excessively long words, as each phoneme takes up memory space. The third term penalizes excessively long representations of the input in terms of the lexicon.

How would this algorithm segment a stream of repeated words such as *dog, dog, dog, …, pig, pig, pig, …*, where each word is repeated N times?[4] Due to the non-linear entropy term, an analytic solution is difficult for the general case. However, we can make certain simplifying assumptions that allow us to derive an analytic solution.

Specifically, I assume (i) that there is a unique optimal segmentation, (ii) that the algorithm does not postulate word boundaries within words and (iii) that the number of words in the lexicon is 2. In this example, the number of unique phonemes is 5. If l_1 and l_2 are the lengths of the words in the lexicon (e.g., the word *pigpigpig* would have a length of 3), the cost function (2) can be rewritten as

$$\begin{aligned} f(S) = \quad & 6 + (\log_2 5)\,(3l_1 + 3l_2) \\ & -\left(\frac{N}{l_1}+\frac{N}{l_2}\right)\sum_{i=1}^{2}\left(\frac{\frac{N}{l_i}}{\frac{N}{l_1}+\frac{N}{l_2}}\right)\log_2\left(\frac{\frac{N}{l_i}}{\frac{N}{l_1}+\frac{N}{l_2}}\right) \end{aligned} \tag{4}$$

Due to the symmetry of the problem, we can assume that $l_1 = l_2 \equiv l$; the cost function (2) can thus be further simplified to

$$f(S) = 6 + 6l\log_2 5 + \frac{2N}{l}. \tag{5}$$

The minimum of this function is for $l = \sqrt{N/(3\log_2 5)}$. For few repetitions of each item, the items in the lexicon are thus the actual words, but as the familiarization sequence gets longer, longer words end up in the lexicon (e.g., *dogdogdog…*). For example, if each pronunciation of a syllable takes roughly 500 ms, there are 60 repetitions per word and minute of a familiarization sequence. Learners extract 3-words-long units after a 1 min familiarization, 4-words-long-units after a 2 min familiarization, 5-words-long units after a 3 min familiarization and so forth.

Appendix B

A similarity-based model of Xu and Tenenbaum (2007b)

Xu and Tenenbaum's (2007b) results can be explained by a simple hypothesis: participants might choose a category that is (i) consistent with the examples they have seen, and (ii) where the examples are most similar to the other items in the category.

The first assumption prevents learners from considering, say, the meaning "Dalmatian" when presented with non-Dalmatian dogs, because the label would not fit the examples. To illustrate the second assumption, I constructed a simple similarity score, using the number of shared category levels of two items as a proxy of their similarity (i.e., subordinate, basic and superordinate). That is, two Dalmatians have a similarity score of 3 (because they share all 3 levels), a Dalmatian and a non-Dalmatian dog have a similarity score of 2, and a Dalmatian and a cat have a similarity score of 1. This similarity score reflects the intuition that items from the same subordinate category (e.g., two poodles) tend to be more similar than two items from the same basic-level category (e.g., a poodle and a Labrador), which, in turn, tend to be more similar than two items from the same superordinate category (e.g., a poodle and a bear).

As shown in Figure B1, this simple model predicts Xu and Tenenbaum's (2007b) results, without using the size principle at all. When the examples are consistent with a subordinate-level category, the similarity is highest for items within such a category; and when they are consistent only with a basic-level category, the similarity is highest in such a category as well.

Appendix C

A function-based version of Gweon et al.'s (2010) model

Gweon et al.'s (2010) model critically depends on the assumption that infants estimate the likelihood of observing squeaking *blue* balls. However, they might well be interested in the balls' squeakiness, irrespective of their color. If so, Gweon et al.'s (2010) model predicts the opposite behavior.

Given a proportion β of blue balls, one can derive different likelihoods for the teacher picking three squeaky blue balls. These likelihoods are presented in Figure C1. The two middle columns show the likelihoods from Gweon et al.'s (2010) model, assuming that infants desire *blue* squeaky balls. The two right-most columns show the likelihoods as-

[4] We need at least two words for obtaining a non-zero entropy.

Examples		Inference: "fep means"	Similarity to other dalmatians (N=2)	non-dalmatian dogs (N=2)	non-dog animals (N=4)	average
Dalmatians		Dalmatian	3			3
		Dog	3	2		2.5
		Animal	3	2	1	1.75
3 Dogs	For dalmatian (N=1)	Dog	3	2		2.5
		Animal	3	2	1	1.75
	For other dogs (N=2)	Dog	2	2		2
		Animal	2	2	1	1.5
	Average	Dog	2.333333333	2		2.166666667
		Animal	2.333333333	2	1	1.583333333

Figure B1. Similarity between training exemplars and other category members in Xu and Tenenbaum's (2007b) experiments. Here, I use the number of category-levels that two items share as a proxy of similarity.

suming that infants care about squeakiness but not about color. The likelihood of picking n squeaky blues balls can be obtained by averaging across the choice strategies of the teacher (see Gweon et al., 2010, for a justification of this average).

Gweon et al. (2010) compared the ratio of (i) the likelihood of the teacher picking three squeaky blue balls if all balls are squeaky and (ii) the likelihood of the teacher picking three squeaky blue balls when only the blues one are squeaky. As mentioned above, this ratio is given by:

$$L = \frac{2\beta^n}{1+\beta^n}. \quad (6)$$

(Compared to Equation (1), and following Gweon et al. (2010), I set α to .5) When 75% of the balls are blue, and the teacher picks 3 balls, the ratio is .59; when 25% of the balls are blue, the ratio is .03. Hence, leaving aside the fact that the model predicts that infants should never squeeze yellow balls, they should be more likely to squeeze yellow balls when 75% of the balls are blue.

The alternative model, where infants care only about the squeakiness of the balls but not their color, reverses the predictions. As can be seen from Figure C1, the corresponding likelihood ratio is (with α set to 0.5):

$$L' = \frac{2}{1+\beta^n}. \quad (7)$$

Keeping Gweon et al.'s (2010) α parameter as a variable, the likelihood ratio is given by:

$$L'_\alpha = \frac{1}{\alpha+(1-\alpha)\beta^n}. \quad (8)$$

It is easy to see that $L'_\alpha = 1$ for $\alpha = 1$ or $\beta = 1$, and $L'_\alpha > 1$ for $\alpha < 1$ and $\beta < 1$. Hence, this model accounts for the fact that infants have a tendency to squeeze balls irrespective of their color.[5]

However, for blue majority populations (with 75% of blue balls), the ratio is 1.41, while it is 1.97 for yellow majority populations (with 25% of blue balls). Hence, the alternative model (incorrectly) predicts that infants should be more likely to squeeze yellow balls when 25% of the balls are blue, though it does account for the fact that infant squeeze the yellow balls to begin with.

[5] In contrast to Gweon et al.'s (2010) model, the partial derivative $\partial_n L'_\alpha = \frac{-(1-\alpha)\beta^n \ln\beta}{(\alpha+(1-\alpha)\beta^n)^2}$ is strictly positive for $\beta \in]0,1[$ and $\alpha \neq 1$. Further, $\lim_{n\to\infty} L'_\alpha = 1/\alpha > 1$.

Squeaky balls	Teacher chooses among	Infants want *blue* squeaky balls: Explanation	Infants want *blue* squeaky balls: *P*	Infants want squeaky balls: Explanation	Infants want squeaky balls: *P*
only blue	squeaky	The hypothesis holds that only blue balls are squeaky, and that the teacher will only sample from these balls. Hence, according to this hypothesis, the teacher will choose squeaky blue balls with probablity 1.	1	The hypothesis holds that the teacher will only sample from squeaky balls. Hence, according to this hypothesis, the teacher will choose squeaky balls with probablity 1.	1
only blue	all	Since the teacher randomly picks balls, she has, for each ball, a chance β for picking a blue/squeaky ball.	β^n	Since the teacher randomly picks balls, she has, for each ball, a chance β for picking a blue/squeaky ball.	β^n
only blue	average		$(1 + \beta^n)/2$		$(1 + \beta^n)/2$
all	squeaky	While all balls are squeaky, the teacher has a chance of β to pick a blue ball.	β^n	Since all balls are squeaky, the probability of picking n squeaky balls is 1.	1
all	all	While all balls are squeaky, the teacher has a chance of β to pick a blue ball.	β^n	Since all balls are squeaky, the probability of picking n squeaky balls is 1.	1
all	average		β^n		1

Figure C1. Likelihoods of the teacher picking n balls of interest out of a box with a proportion of β blue balls. The middle two columns present Gweon et al.'s (2010) model, in which infants seek blue squeaky balls. The likelihood ratio in favor of the hypothesis that all balls are squeaky is given by $2\beta^n/(1 + \beta^n)$. The rightmost columns present an alternative model, where infants just care about the squeakiness of the balls, irrespective of their color. In that case, the likelihood ratio is $2/(1 + \beta^n)$.