

Humans store ~ 1.5 megabytes during language acquisition:
information theoretic bounds

Francis Mollica^{1†} & Steven T. Piantadosi²

¹Department of Brain & Cognitive Sciences, University of Rochester

²Department of Psychology, University of California, Berkeley

†mollicaf@gmail.com

keywords: linguistic memory, information theory, Fermi estimation

Abstract

We introduce theory-neutral estimates of the amount of information learners know about how language works in order to quantify the capacities of the inferential mechanisms that support language learning. We provide estimates at several levels of English linguistic analysis: phonemes, wordforms, lexical semantics, word frequency and syntax. Our best guess estimate for human knowledge of language is 12.5 million bits, the majority of which is lexical semantics. Interestingly, very little of this information is syntactic, even in our upper bound analyses. Generally, our results indicate the learners must extract, on average, nearly 2000 bits of information about how their language works *each day* of their early lives.

1 Introduction

One of the foundational debates about human language centers on an issue of scale: is the amount of information about language that is learned substantial (empiricism) or minimal (nativism)? Despite theoretical debates on the how much of language is or can be learned, the general question of the amount of information that *must* be learned has yet to be quantified. Here, we provide an estimate of the amount of information learners must extract in the process of language acquisition. We provide *lower-bound*, *best guess*, and *upper-bound* estimates of this information, using a “back of the envelope” approach that is popular in physics. During the testing of the atomic bomb, physicist Enrico Fermi famously estimated the strength of the bomb by dropping scraps of paper as the blast passed. He noted that they were blown about 8 feet by the explosion and announced out loud that the blast was equivalent to about 10,000 tons of TNT [1]—the true answer was 18,000 tons, meaning Fermi’s crude experiment and quick calculation got the right answer to within a factor of two. Similar back-of-the-envelope *Fermi*-calculations are commonly used in physics as a sanity check on theories and computations.¹ However, such sanity checks are needed—although rarely applied—in fields that suffer from under-constrained theories, like psychology.

We apply this approach of rough estimation in order to quantify a lower-bound on the number of *bits per day* that language learners must extract and remember from their environments. While a substantial literature has focused on the differences between nativist and empiricist approaches, when they are translated into the domain of information theory, nativism and empiricism may blur together. Specifically, a constrained nativist space of possibilities with one set of learning biases may require the same number of bits of information as an unconstrained, large space of hypotheses with another set of biases. Since we do not know much about which language learning biases children have, theories about how the hypotheses are constrained (or not) do not unambiguously determine the number of bits learners must store.

So, instead of debating nativism versus empiricism, we take up the challenge of quantifying how much information of language must be learned in order to inform the underlying acquisition theories. To avoid dependence on a particular representation scheme, we focus on the possible *outcomes* of learning a language, i.e., we compute the number of bits required to specify the target outcome out of a plausible space of logically possible alternatives. To avoid dependence on a particular learning algorithm, we study the problem abstractly without reference to *how* learning works, but rather based on how much a relatively unbiased (e.g. maximum entropy) learner would have to store.

Our study is inspired by prior work which has aimed to characterize the capacity of human memory. Early on, memory capacity was approached from a neuroanatomical perspective. Upper bounds on memory capacity have been estimated via the number of synapses in cortex (10^{13} bits) or the number of impulses conducted in the brain across a lifetime (10^{20} bits) [2]. More recently, bounds for information transfer in neural coding have been estimated using information theoretic techniques [3]. Working from behavioral performance, Landauer [4] used a variety of techniques to estimate the number of bits of information humans must have in memory in order to show a given level of task performance. In one example, he converted accuracy in a recognition memory task to bits by imagining that each image was stored using a random code. This technique has been used recently by [5] in a large-scale empirical study, which estimated that human memory could encode $2^{13.8}$ unique items. Even more recently, [6] estimated 4 and 6 year old children’s memory capacity to be $2^{10.43}$ unique items. Strikingly, both of these estimates lie within Landauer’s estimated range of 10-14 bits per item. This study also used a dictionary study to estimate the number of words that Stanford students knew, and converted the estimates for a phonetic code into bits, requiring about 30 – 50 bits per word [4]. All of his estimates converged on the same order of magnitude, suggesting that the “functional capacity” for human memory is on the order of 10^9 bits. A detailed critique of Landauer can be found in [7], with a response given by [8].

Our focus here is on estimating *linguistic* knowledge across multiple levels of structure and function: phonemes, wordforms, lexical semantics, word frequency and syntax. At each level, there is a large space of logically possible linguistic representations (e.g., acoustic cue values, syntactic parses). The challenge for learners is to discover and store which representations are used in their language. Tools in information theory allow us to estimate the relevant quantities. First, we assume that before learning, children begin

¹These computations are also used as a training exercise that allows surprising quantities to be approximated. An example is to compute the thickness of a car tire that is worn off with each rotation. Here’s a hint: you can use your knowledge of how many miles car tires last for and how much thickness they lose over their lifetime.

with a certain amount of uncertainty over the required representation, R , denoted $\mathbf{H}[R]$. Shannon entropy [9] quantifies the number of bits that must be received on average to remove uncertainty about what R is the true one. After observing some data D , learners will have a new amount of uncertainty (perhaps zero) over R , denoted $\mathbf{H}[R | D]$. Note that here, D is not a random variable, but rather a specific value of data in learning a given language.

We can formalize the amount of information that D provides about R , here denoted $\Delta\mathbf{H}[R | D]$ as the difference between $\mathbf{H}[R]$ and $\mathbf{H}[R | D]$:

$$\Delta\mathbf{H}[R | D] = \mathbf{H}[R] - \mathbf{H}[R | D] = - \sum_{r \in R} P(r) \log P(r) + \sum_{r \in R} P(r | D) \log P(r | D). \quad (1)$$

This quantity, i.e. the reduction in entropy, gives the amount of information that D (e.g. data from learning) provides about a representation R .² Thus, in order to estimate the amount of information learners must have acquired, it suffices to estimate their uncertainty before learning, $\mathbf{H}[R]$, and subtract from it their uncertainty after learning $\mathbf{H}[R | D]$. The resulting value will tell us the number of bits of information that the learning data D has provided.³

2 Results

Table 1: Summary of estimated bounds across levels of linguistic analysis

Section	Domain	Lower bound	Best Guess	Upper bound
2.1	Phonemes	375	750	1500
2.2	Phonemic Wordforms	200,000	400,000	640,000
2.3	Lexical Semantics	553,809	12,000,000	40,000,000
2.4	Word Frequency	40,000	80,000	120,000
2.5	Syntax	134	697	1394
Total (bits)		794,318	12,481,447	40,762,894
Total per day (bits)⁴		121	1,900	6204

We will build up our estimates separately for each linguistic domain. The results of each section are summarized in Table 1. Supplementary Table S1 summarizes the key assumptions behind each of our estimations.

2.1 Information about Phonemes

Our phonemic knowledge enables us to perceive discrete linguistically-relevant sounds, or phonemes, from rich high-dimensional but noisy speech signals. Before a child knows the sounds of their language, they have uncertainty over the acoustic features of speech sounds. After learning their language, children have much less uncertainty over the acoustic features of speech sounds as they now have several acoustic cues to help them identify phonemes. Following the above logic, the decrease in the amount of uncertainty children have about where their speech sounds lie in acoustic space is the amount of information they must now store about phoneme cues.

Identifying linguistically relevant acoustic cues has proven challenging for scientists, as there is no obvious invariance, or uniquely identifying component, in speech sounds. For our estimation, we analyze the information stored for three well studied acoustic cues: voice onset time (VOT) in ms—a cue to voiced-voiceless distinctions (e.g., the difference between /p/ and /b/), central frication frequency in barks—a cue to the place of articulation for fricatives (e.g., the difference between /f/ and /s/), and the first two formant frequencies of vowels in mels—a cue to vowel identity. We assume that initially learners have maximum

²The average of $\Delta\mathbf{H}[R | D]$ over D is the *mutual information* between R and D [10].

³In the case of continuous distributions, (1) has continuous analogs where the sums turn into integrals.

⁴For this value, we assume language is learned in 18 years of 365 days.

uncertainty along each cue R , following uniform distributions bounded by the limits of perception. In this case, each $r \in R$ has an equal probability of $P(r) = 1/(B - A)$, giving

$$\mathbf{H}[R] = - \int P(r) \log P(r) dr = \log(B - A), \quad (2)$$

where B and A are respectively the upper and lower bounds of perception. For VOT, we assume the range is -200 to 200 ms. For frequencies, we assume bounds on human hearing of 20 to $20,000$ Hz, which translate to $0.2 - 24.6$ in barks and $32 - 3817$ in mels. As a measure of the uncertainty over the cue dimension after learning, we will assume that speaker’s knowledge is captured by a normal prior distribution, giving $\mathbf{H}[R | D]$ as

$$\mathbf{H}[R | D] = \int N(x | \mu, \sigma) \log N(x | \mu, \sigma) dx = \frac{1}{2} \log(2\pi e \sigma^2) \quad (3)$$

where N is a normal probability density function,⁵ and μ and σ are the participants’ inferred mean and standard deviation. To find σ for real humans, we use the values inferred by [11, Table 7] to account for the perceptual magnet effect.⁶

We find that language users store 3 bits of information for voiceless VOT, 5 bits for voiced VOT, 3 bits for central frication frequency and 15 bits for formant frequencies. As these acoustic cues are only a subset of the cues required to identify consonant phonemes, we assume that consonants require three cues; each cue requiring 5 bits of information. For vowels, we do not adjust the 15 bits of information conveyed by formant frequencies. As a best guess, again paying attention to primarily the order of magnitude, we assume there are 50 phonemes each requiring 15 bits, totaling 750 bits of information. For lower and upper estimates, we introduce a factor of two error [375-1500 bits].

2.2 Information about Wordforms

When dealing with wordforms, the first challenge is to define a “word.” The term “word” is ambiguous and could mean lemmas, phonological forms, families of words, etc. Entire dissertations could be (and have been) written on these distinctions. These difficulties are in part why the Fermi-approach is so useful: we don’t need to make strong theoretical commitments in order to study the problem if we focus on rough estimation of orders of magnitude. Estimates of the number of words children acquire range on the order of 20,000-80,000 total wordforms [12]. However, when words are grouped into families (e.g. “dog” and “dogs” are not counted separately) the number known by a typical college student is more in the range of 12,000 – 17,000 [13, 14]; although see [15] for an estimate over twice that size. Lexical knowledge extends beyond words too. [16] estimates that the average adults understands 25,000 idioms, items out of the view of most vocabulary studies. Our estimates of *capacity* could of course be based on upper-bounds on what people *could* learn, which, to our knowledge, have not been found. Looking generally at these varied numbers, we’ll use an estimate of 40,000 as the number of essentially unique words/idioms in a typical lexicon.

The most basic thing each learner must acquire about a word is its phonemic wordform, meaning the sequence of phonemes that make up its phonetic realization. If we assume that word forms are essentially memorized, then the entropy $\mathbf{H}[R | D]$ is zero after learning—e.g. for all or almost all words, learners have no uncertainty of the form of the word once it has been learned. The challenge then is to estimate what $\mathbf{H}[R]$ is: before learning anything, what uncertainty should learners have? To answer this, we can note that $\mathbf{H}[R]$ in (1) can be viewed as an *average* of the negative log probability of a wordform, or $-\log P(R)$. Here, we use a language model to estimate the average negative log probability of the letter sequences that make up words and view this as an estimate of the amount of entropy that has been removed for *each* word. In other words, the average surprisal of a word under a language model provides one way to estimate the amount of uncertainty that learners who know a given word must have removed.^{7,8}

⁵Using a normal distribution with the domain truncated to our perceptual bounds does not change our estimate.

⁶For vowels, we extend these distributions to their multidimensional counterparts as formant space is (at least) two dimensional.

⁷In this view, we neglect the complexity of the language model, which should be a much smaller order of magnitude than the number of bits required for the entire lexicon.

⁸Analogously, we can view the surprisal as the number of bits that must be remembered or encoded for a particular outcome—e.g. to learn a specific wordform.

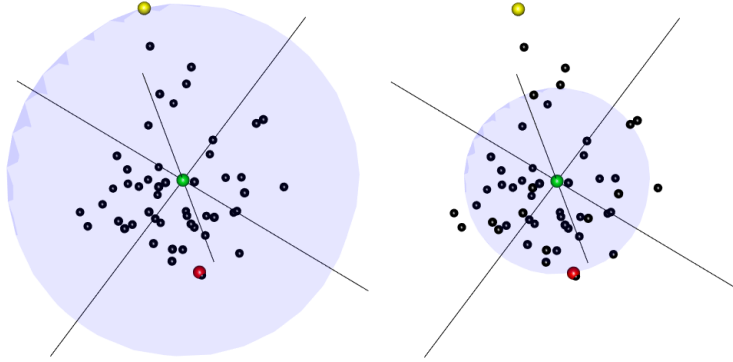


Figure 1: The shaded spheres represent uncertainty in semantic space centered around a word (in green). Left: The uncertainty is given with respect to the word’s farthest connection in semantic space (in yellow), representing R . Right: The uncertainty is given with respect to the N^{th} nearest neighbor of the word (in red), representing r . The reduction in uncertainty from R to r reflects the amount of semantic information conveyed by the green word.

To estimate these surprisals, we used the CELEX database [17], we computed the surprisal of each word under 1-phone, 2-phone, 3-phone and 4-phone models [see 18] trained on the lexicon. This analysis revealed that 43 bits per word on average are required under the 1-phone, 33 bits per word under the 2-phone, 24 under the 3-phone and 16 under the 4-phone model. Noting the sharply decreasing trend, we will assume a lower bound of about 5 bits per word to store the phonetic sequence, a “best guess” of 10 bits per word and an upper bound of 16 as in the 4-phone.⁹ When our best guess is multiplied by the size of the lexicon (40,000 words), that gives an estimate of 400,000 [200,000 – 640,000] bits of lexical knowledge about the phonetic sequences in words.

2.3 Information about Lexical Semantics

The information contained in lexical semantics is difficult to evaluate because there are no accepted theories of semantic content, or conceptual content more generally [19]. However, following Fermi, we can make very simplified assumptions and try to estimate the general magnitude of semantic content. One way to do this is to imagine that the set of word meanings are distributions in a N -dimensional semantic space. If we assume that the entire space is a Gaussian with standard deviation R and the standard deviation of an individual word meaning is r , then we can compute the information contained in a word meaning as the difference in uncertainty between a N -dimensional Gaussian with radius R as compared to one with radius r . The general logic is shown in Figure 1. The “space” shown here represents the space of semantic meanings, and words are viewed as small distributions in this space covering the set of things in the extension of the word’s meaning. Thus, when a learner acquires a word like “accordion”, they know that the word refers to some relatively small subset (of size r) of possible objects, but they may not be certain on the details (does the extension cover harmoniums? Concertinas? Bayans?). The reduction in entropy from a total semantic space of size R —no idea what a word means—to one of size r is what we use to approximate the amount of information that has been learned.

Equation (3) above gives the change in entropy for a one-dimensional Gaussian. However, the dimensionality of semantic space is considerably larger. In the case of an N -dimensional Gaussian, with independent dimensions and (constant, or homogeneous) standard deviation σ in each dimension, the entropy is given by:

$$\mathbf{H}[R] = \frac{N}{2}(1 + \log 2\pi + \log \sigma). \quad (4)$$

This means that if we go from a R standard deviation Gaussian to a r standard deviation one, the amount

⁹As the average word length in this database is ~ 7.5 phonemes, this gives slightly over one bit per phoneme.

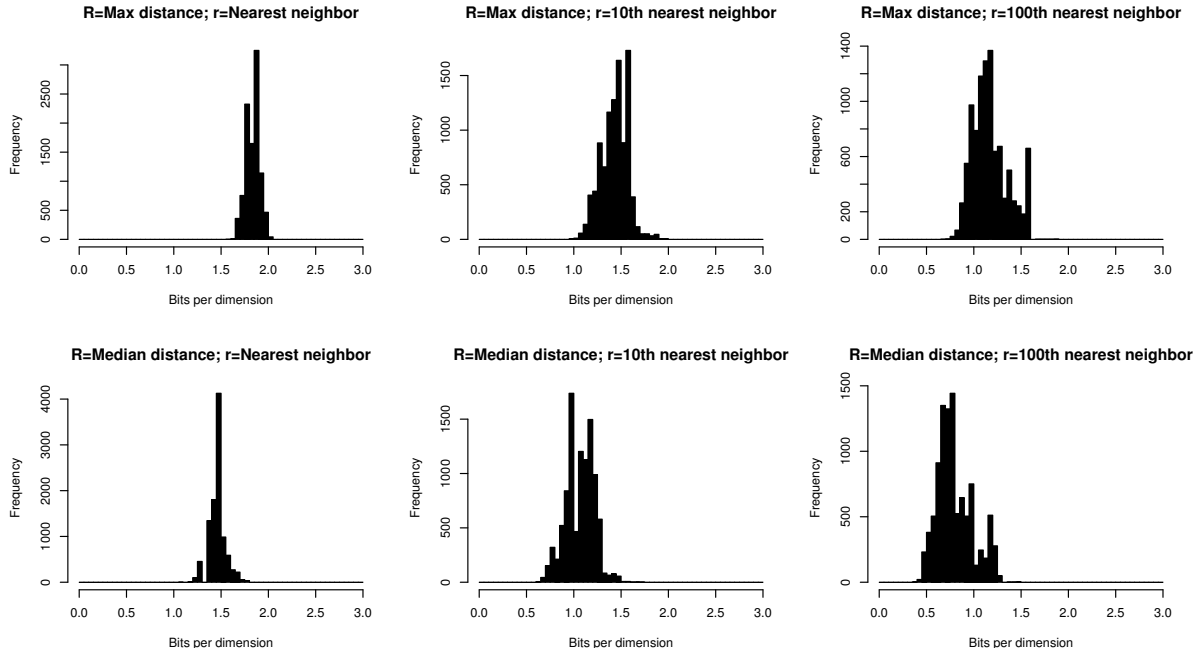


Figure 2: Histograms showing the number of bits-per-dimension ($\frac{1}{2} \log \frac{R}{r}$) for various estimates of R and r . These robustly show that 0.5 – 2.0 bits are required to capture semantic distances.

of information we have learned is the difference between these entropies,

$$\Delta \mathbf{H}[R | D] = \frac{N}{2}(1 + \log 2\pi + \log R) - \frac{N}{2}(1 + \log 2\pi + \log r) = \frac{N}{2} \log \frac{R}{r} \quad (5)$$

We estimate R and r in several different ways by looking at WordNet [20] to determine the closeness of each word to its neighbors in semantic space. In particular, we take r to be a characteristic distance to nearby neighbors (e.g. the closest neighbors), and R to be a characteristic distance to far away ones (e.g. the max distance). Note, that this assumes that the size of a Gaussian for a word is about the same size as its distance to a neighbor, and in reality this may *under-estimate* the information a word meaning contains because words could be much more precise than their closest semantic neighbor.

Figure 2 shows $\frac{1}{2} \log \frac{R}{r}$ for several estimates of R and r for 10,000 random nouns in WordNet. The likely values fall within the range of 0.5 – 2.0 bits. Because we are plotting $\frac{1}{2} \log \frac{R}{r}$ and not $\frac{N}{2} \log \frac{R}{r}$, these values may be interpreted as the number of *bits per dimension* that lexical semantics requires. For instance, if semantic space was one-dimensional then it would require 0.5 – 2.0 bits per word; if semantic space were 100-dimensional, lexical semantics would require 50 – 200 bits per word. The nearness of these values to 1 means that even continuous semantic dimensions can be viewed as *approximately* binary in terms of the amount of information they provide about meaning.

The dimensionality of semantic space has been studied by [21] and [22], with numbers ranging from 100 – 500 dimensions. Our best guess will use 1 bit per dimension and 300 dimensions following [21] for 12,000,000 bits. Our upper bound uses 2 bits-per-dimension and 500 dimensions for a total of 40,000,000 bits.

For our lower-bound in this domain, we may pursue a completely alternative technique to compute the lower-bound—one which, surprisingly, gives a similar order of magnitude as our best guess. If there are 40,000 lexical items that must be learned, we can assume that they correspond to 40,000 distinct concepts (a la *principle of contrast* [23]). In the “most nativist” case, favored by Fodor [24], we could assume that there are a corresponding 40,000 meanings for these words that learners innately have. In this case, the problem of learning is figuring out which of the 40000! pairings of words and concepts is the correct one. It will take $\log_2(40000!) \approx 553,809$ bits of information to specify which of these is correct. We will use this as

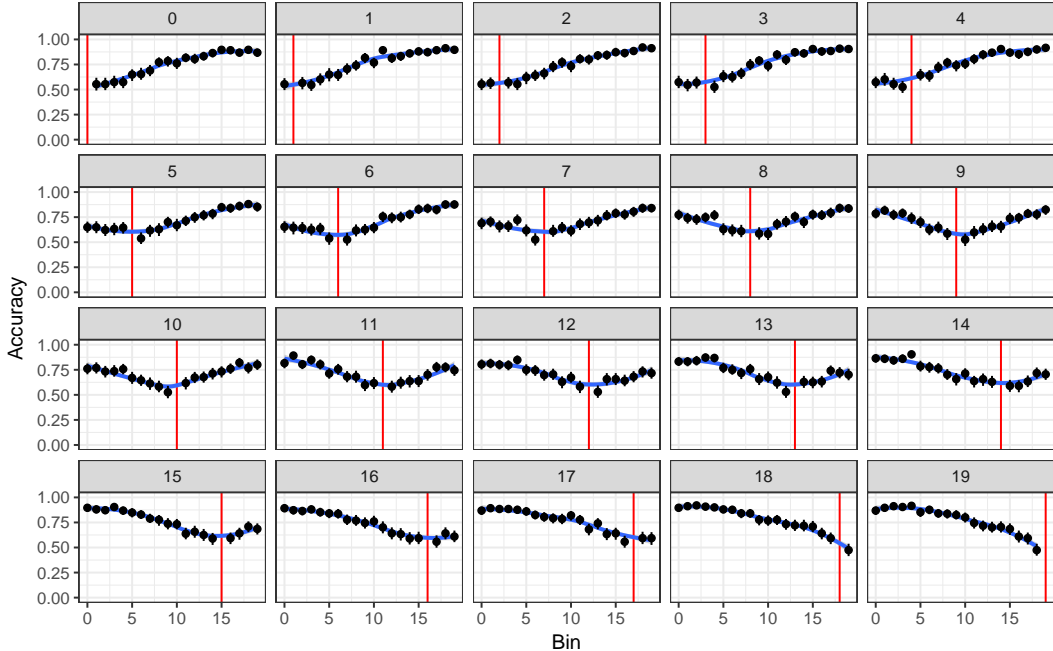


Figure 3: Accuracy in frequency discrimination accuracy as a function of log word frequency bin faceted by log reference word frequency bin. Vertical red lines denote within bin comparison. Line ranges reflect 95% bootstrapped confidence intervals.

our lower-bound. While this seems like an unmanageable task for the child, it’s useful to imagine how much information is conveyed by a single pedagogical learning instance. Our estimate is derived by a combinatorial argument: to choose the first word’s meaning, there are N choices, the second there $N - 1$, and so on. The total number of choices is therefore $N \cdots (N - 1) \cdots (N - 2) \cdots (N - 40000) = N! / (N - 40000)!$. So if initially $N = 400000$ (553,809 bits), there will be $N = 39999$ (553,794 bits) after one correct mapping is learned, meaning that a single pedagogical instance rules out 40,000 possible pairings or, equivalently conveys 15.29 bits.

2.4 Information about Word Frequency

Word frequencies are commonly studied in psychology as factors influencing language processing and acquisition [e.g., 25, 26, 27, 28, 29, 30] as well as for their peculiar Zipfian distribution [31]. However, relatively little work has examined the fidelity of people’s representation of word frequency, which is what is required in order to estimate how much people know about them. In one extreme, language users might store perhaps only a single bit about word frequency, essentially allowing them to categorize high vs. low frequency words along a median split. On the other extreme, language users may store information about word frequency with higher fidelity—for instance, 10 bits would allow them to distinguish 2^{10} distinct levels of word frequency. Or, perhaps language learners store a full ranking of all 40,000 words in terms of frequency, requiring $\log(40000!) = 553,809$ bits of information.

In an experimental study, we asked participants from Mechanical Turk ($N = 251$) to make a two-alternative forced choice to decide which of two words had higher frequency¹⁰. Words were sampled from the lexical database SUBTLEX [32] in 20 bins of varying log frequency. We removed words below the bottom 30th percentile (frequency count of 1) and words above the upper 99th percentile in word frequency in order to study the intermediate-frequency majority of the lexicon. Each participant completed 190 trials.

Participants’ accuracy in answering is shown in Figure 3. The i ’th subplot shows participants’ accuracy

¹⁰Participants were instructed that we were interested in their first impression and that there was no need to look up word frequencies.

(y -axis) in distinguishing the i 'th bin from each other j 'th bin, with the red line corresponding to $i = j$. This shows, for instance, that people are poor at distinguishing very close i and j (near the red line), as should be expected.

Participants' overall accuracy in this task was 76.6%. Neglecting the relatively small difference in accuracy (and thus fidelity) with a words' absolute frequency, this accuracy can be modeled by imagining that participants store M levels of word frequencies. Their error rate on this task will then be given by the probability that two words fall into the same bin, or $1/M$. Setting $1/M = 1 - .766$ gives $M \approx 4$, meaning that participants appear to track approximately four categories of frequencies (e.g. high, medium-high, medium-low, low). This trend can also be observed in Figure 3, where the flat bottom of the trough in each plot is approximately 5 bins wide, meaning that each bin cannot be well distinguished from its 5 nearest neighbors (giving a total effective number of bins for participants as $20/5 = 4$).

If $M = 4$, then participants would only need to learn $\log 4 = 2$ bits of information about a word's frequency, as a best guess. This would yield a total of $2 \cdot 40,000 = 80,000$ bits total across the entire lexicon. We construct our lower- and upper- bounds by introducing a factor of two error on the computation (e.g. per word lower bound is 1 bit and upper is 3 bits). It is important to note that by assuming objective frequency rankings, our estimate is conservative. If we could analyze participants' responses with regard to their subjective frequency rankings, we would expect to see greater accuracy reflecting higher resolution representations of frequency.

2.5 Information about Syntax

Syntax has traditionally been the battle ground for debates about how much information is built in versus learned. Indeed, syntactic theories span the gamut from those that formalize a few dozen binary parameters [33, 34] to ones that consider alternative spaces of infinite models [e.g. 35, 36] or data-driven discovery from the set of all parse trees [37]. In the face of massively incompatible and experimentally under-determined syntactic theories, we aim here to study the question in a way that is as independent as possible from the specific syntactic formalism.

We do this by noting that syntax provides information to parse every sentence of English. In many cases, the sentences of English will share syntactic structure. However, we can imagine a set s_1, s_2, \dots, s_n of sentences which share as little syntactic structure as possible between each s_i and s_j . For instance,

$$\text{Bill [met John]} \quad \text{and} \quad \text{[Jill's sister] cried} \quad (6)$$

both have three words but have largely non-overlapping syntactic structures due to the use of a transitive verb in the first and a possessive and intransitive verb in the second. We will call these "essentially independent" sentences when they share almost no syntactic structure. In this case, the bits specifying these parses can be added together to estimate the total information learners know. If the sentences were not essentially independent in terms of syntactic structure, information from one sentence would tell us how to parse information from another, and so adding together the information for each would be an over-estimate over learners' knowledge.

We assume that learners who do not know anything about a parse of a sentence s_i start with a maximum entropy distribution over each parse, assigning each an equal probability of one over the number of logically possible parses of s_i , so that

$$\mathbf{H}[R] = - \sum_{r \in R} \frac{1}{\#parses} \log \frac{1}{\#parses} = \log(\#parses). \quad (7)$$

We assume the knowledge of an adult leaves zero uncertainty in general, yielding $\mathbf{H}[R | D] = 0$ so that

$$\Delta \mathbf{H}[R | D] = \mathbf{H}[R] - \mathbf{H}[R | D] = \log(\#parses) \quad (8)$$

for a single sentence s_i . In general, the number of logically possible parses can be computed as the number of binary trees over s_i , which is determined only by the length of s_i . The $(l - 1)$ 'th Catalan number gives the number of possible binary parses for a sentence of length l . Then, the number of bits required to specify *which* of these parses is correct is given by $\log C_{l-1}$. The Catalan numbers are defined by

$$C_n = \frac{1}{n+1} \binom{2n}{n}. \quad (9)$$

As an example, to determine each of (6), knowledge of syntax would have to specify $\log C_2 = 1$ bit, essentially specifying whether the structure is $((\circ \circ) \circ)$ or $(\circ (\circ \circ))$. But C_n grows exponentially—for instance, $C_{10} = 16796$, requiring 9.7 bits to specify which parse is correct for an 11-word sentence.

Looking at a collection of sentences, if s_i has length $l(s_i)$, then the total amount of information provided by syntax will be given by

$$\sum_{i=1}^n \log C_{(l(s_i)-1)}. \tag{10}$$

Again, (10) assumes that there is no syntactic structure shared between the s_i —otherwise (10) over-estimates the information by failing to take into account the fact that some bits of information about syntax will inform the parses of distinct sentences. Our upper and lower bounds will take into account uncertainty about the number of distinct sentences s_i that can be found.

To estimate the number of such sentences, we use the textbook linguistic examples studied by [38]. They present 111 sentences that are meant to span the range of interesting linguistic phenomena and were presented independently in [39]. Our best estimate is therefore (10) taking s_i to be the lengths of these sentences. We take the lower-bound to be (10) where $l(s_i)$ is *half* the sentence length of s_i , meaning that we assume that only half of the words in the sentence participate in a structure that is independent from other sentences. For an upper bound, we consider the possibility that the sentences in [38] may not cover the majority of syntactic structures, particularly when compared to more exhaustive grammars like [40]. The upper bound is constructed by imagining that linguists could perhaps construct two times as many sentences with unique structures, meaning that we should double our best guess estimate. Notably, these tactics to bound the estimate do not qualitatively change its size: human language requires very little information about syntax, 697 [134 – 1394] bits. In either case, the number is much smaller than most other domains.

3 Discussion

Summing across our estimates for the amount of information language users store about phonemes, word-forms, lexical semantics, word frequency and syntax, our best guess and upper bound are on the order of 10 million bits of information, the same order as [4]’s estimate for language knowledge. It may seem surprising but, in terms of digital media storage, our knowledge of language fits compactly on a floppy disk. The best-guess estimate implies that learners must be remembering 1000-2000 bits *per day* about their native language, which is a remarkable feat of cognition. Our lower bound is around a million bits, which implies that learners would remember around 120 bits each day from birth to 18 years. To put our lower estimate in perspective, each day for 18 years a child must wake up and remember, perfectly and for the rest of their life, an amount of information equivalent to the information in this sequence,

```
0110100001101001011001000110010001100101011011100110000101100011
01100011011011110111001001100100011010010110111101101110
```

Naturally, the information will be encoded in a different format—presumably one which is more amenable to the working of human memory. But in our view, both the lower and best-guess bounds are explainable only under the assumption that language is grounded in remarkably sophisticated memory mechanisms.

There are several limitations to our methods, which is part of the reason we focus on orders of magnitude rather than precise estimates. First, our estimates are rough and only hold under our simplifying assumptions (listed in Table S1). Second, there are several domains of linguistic knowledge whose information content we do not estimate here including word predictability, pragmatic knowledge, knowledge of discourse relations, prosodic knowledge, models of individual speakers and accents, among others. Many of these domains are difficult because the spaces of underlying representations are not sufficiently well formalized to compute information gains (e.g. in pragmatics or discourse relations). In other areas like people’s knowledge of probable sequences of words, the information required is difficult to estimate because the same content can be shared between constructions or domains of knowledge (e.g. the knowledge that “Mary walks” and “John walks” are high probability may not be independent from each other, or from knowledge about the lexical semantics of the verb). We leave the estimation of the amount of information language users store about these domains of language to further research.

Importantly, our estimates vary on orders of magnitude across levels of representation. These differences could suggest fundamental differences in the learning mechanism for specific language learning problems. As these analyses show, the majority of information humans store about language is linked to words, specifically lexical semantics, as opposed to other systems of language knowledge, such as phonemes and syntax. In fact, the estimate for syntax is of a similar order of magnitude proposed by some nativist accounts, in that the number of bits required for syntax is in the hundreds, not tens of thousands or millions. To illustrate, if syntax learning is principally completed in the first 5 years, children would have to learn a single bit about syntax every 2-3 days on average. Despite this, the possible outcomes for learners in our best guess for syntax consists of $2^{697} \approx 10^{210}$ different systems. In other words, learners still would need the ability to navigate an immense space of possibilities, far greater than the number of atoms in the universe ($\sim 10^{80}$). In the other areas of language, even more enormous hypothesis spaces are faced as well, pointing to the existence of powerful inferential mechanisms.

Turning back to nativism and empiricism, it is interesting that the majority of the learnability debates have primarily centered on syntactic development, which in terms of bits, is way less information to learn than that of even just a few word meanings. Perhaps, it is an accident that so much has played out in syntax where the learning problem is comparably easier to formalize; and yet at the same time it is unclear if the proposed models would even be identifiable given the acquisition data we could observe. Based on our estimates, the real learning challenge appears to be lexical semantics, which to our knowledge, has no nativist learning accounts (c.f. [24]). Any innate human capacity for language learning is most likely to be for learning lexical semantics.

4 Ethics

This project was approved by the University of Rochester RSRB

5 Data Accessibility Statement

All data and code can be found in [41].

6 Author Contributions

The authors contributed equally to the design, coding, and writing of this manuscript.

7 Competing Interests

We have no competing interests.

8 Acknowledgements

The authors would like to thank Daniel Gildea for helpful discussions and Evelina Fedorenko for comments on early drafts.

References

- [1] David N Schwartz. *The Last Man who Knew Everything: The Life and Times of Enrico Fermi, Father of the Nuclear Age*. Hachette UK, 2017.
- [2] John Von Neumann. *The Computer and the Brain*. Yale University Press, 1958.
- [3] Alexander Borst and Frédéric E Theunissen. Information theory and neural coding. *Nature neuroscience*, 2(11):947–957, 1999.

- [4] Thomas K Landauer. How much do people remember? some estimates of the quantity of learned information in long-term memory. *Cognitive Science*, 10(4):477–493, 1986.
- [5] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.
- [6] Katrina Ferrara, Sarah Furlong, Soojin Park, and Barbara Landau. Detailed visual memory capacity is present early in childhood. *Open Mind*, 1(3):136–147, 2017.
- [7] Lawrence Hunter. Estimating human cognitive capacities: A response to Landauer. *Cognitive Science*, 12(2):287–291, 1988.
- [8] Thomas K Landauer. An estimate of how much people remember, not of underlying cognitive capacities. *Cognitive Science*, 12(2):293–297, 1988.
- [9] C.E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:623–656, 1948.
- [10] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [11] Yakov Kronrod, Emily Coppess, and Naomi H Feldman. A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, 23(6):1681–1712, 2016.
- [12] Jeremy M Anglin, George A Miller, and Pamela C Wakefield. Vocabulary development: A morphological analysis. *Monographs of the society for research in child development*, pages i–186, 1993.
- [13] Eugene B Zechmeister, Andrea M Chronis, William L Cull, Catherine A D’Anna, and Noreen A Healy. Growth of a functionally important lexicon. *Journal of Literacy Research*, 27(2):201–212, 1995.
- [14] Catherine A D’Anna, Eugene B Zechmeister, and James W Hall. Toward a meaningful definition of vocabulary size. *Journal of Literacy Research*, 23(1):109–122, 1991.
- [15] Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant’s age. *Frontiers in psychology*, 7, 2016.
- [16] Ray Jackendoff. *The architecture of the language faculty*. Number 28. MIT Press, 1997.
- [17] R Baayen, Richard Piepenbrock, and L Gulikers. Celex2 ldc96l14. *Web Download. Philadelphia: Linguistic Data Consortium*, 1995.
- [18] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.
- [19] Stephen Laurence and Eric Margolis. Concepts and cognitive science. *Concepts: core readings*, pages 3–81, 1999.
- [20] Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.
- [21] Thomas K Landauer and Susan T Dumais. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.
- [22] Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78, 2017.
- [23] Eve V Clark. The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition*, 1:33, 1987.

- [24] Jerry A Fodor. *The language of thought*. Harvard University Press, 1975.
- [25] Kenneth I Forster and Susan M Chambers. Lexical access and naming time. *Journal of verbal learning and verbal behavior*, 12(6):627–635, 1973.
- [26] David A Balota, Michael J Cortese, Susan D Sergent-Marshall, Daniel H Spieler, and Melvin J Yap. Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133(2):283, 2004.
- [27] David C Plaut, James L McClelland, Mark S Seidenberg, and Karalyn Patterson. Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological review*, 103(1):56, 1996.
- [28] Marco Zorzi, George Houghton, and Brian Butterworth. Two routes or one in reading aloud? a connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance*, 24(4):1131, 1998.
- [29] Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204, 2001.
- [30] Wayne S Murray and Kenneth I Forster. Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, 111(3):721, 2004.
- [31] Steven T Piantadosi. Zipf’s word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.
- [32] Marc Brysbaert and Boris New. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990, 2009.
- [33] Kenneth Wexler and M Rita Manzini. Parameters and learnability in binding theory. In *Parameter setting*, pages 41–76. Springer, 1987.
- [34] Karen Thompson Kohl. *An analysis of finite parameter learning in linguistic spaces*. PhD thesis, Massachusetts Institute of Technology, 1999.
- [35] Amy Perfors, Joshua B Tenenbaum, and Terry Regier. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338, 2011.
- [36] Nick Chater and Paul Vitányi. ‘ideal learning’ of natural language: Positive results about learning from positive evidence. *Journal of Mathematical psychology*, 51(3):135–163, 2007.
- [37] Rens Bod, Remko Scha, Khalil Sima’an, et al. *Data-oriented parsing*. Citeseer, 2003.
- [38] Jon Sprouse and Diogo Almeida. Assessing the reliability of textbook data in syntax: Adger’s core syntax. *Journal of Linguistics*, 48(03):609–652, 2012.
- [39] David Adger. *Core syntax: A minimalist approach*, volume 33. Oxford University Press Oxford, 2003.
- [40] Rodney Huddleston, Geoffrey K Pullum, et al. The cambridge grammar of english. *Language. Cambridge: Cambridge University Press*, pages 1–23, 2002.
- [41] Francis Mollica and Steven T. Piantadosi. Humans store 1.5 megabytes during language acquisition: information theoretic bounds, Aug 2018. URL osf.io/ga9th.

Table S1: The assumptions we make in our estimates.

Section	Domain	Assumptions
2.1	Phonemes	<ol style="list-style-type: none"> 1. The language system must contain information about acoustic cues to phoneme identity. 2. The maximum entropy over the frequency dimension can be represented as a uniform distribution over audible frequency ranges. 3. The maximum entropy over the VOT dimension can be represented as a uniform distribution ranging from -200 to 200 ms. 4. The variance in language users' representations of acoustic cues for phonemes can be well approximated by normal distributions following [11].
2.2	Phonemic Wordforms	<ol style="list-style-type: none"> 1. The language system favors compression of statistical co-occurrences. 3. The cost of specifying a language model over phonemes is negligible. 4. Adult language users have a lexicon of 40,000 lexical entries. 5. The sample of words we used to induce our estimate is an adequate approximation to the adult lexicon.
2.3	Lexical Semantics	<ol style="list-style-type: none"> 1. Semantic space can be represented as a multivariate normal distribution with independent dimensions. 2. The maximum entropy over the space can be approximated by a normal distribution whose standard deviation is the maximum distance between words. 3. What learners come to know about the semantics of words narrows the distribution over semantic space based on distance to the nearest semantic neighbor. 4. Adult language users have a lexicon of 40,000 lexical entries. 5. Our sample of words is a decent approximation to the distances of the average word.
2.4	Word Frequency	<ol style="list-style-type: none"> 1. Errors in word frequency discrimination are a result of insufficient representational resolution. 2. Subjective frequency rankings are well approximated by objective frequency rankings (via corpus statistics). 3. Adult language users have a lexicon of 40,000 lexical entries. 4. The sample of words we used in our experiment are representative of the words in the adult lexicon.
2.5	Syntax	<ol style="list-style-type: none"> 1. The language system must contain information to uniquely identify one binary parse tree from all possible binary parse trees. 2. The maximum entropy over syntactic parses is given by the number of binary parse trees. 3. Sentences from [38] are a good approximation/coverage of the essentially independent syntactic components of English grammar.