Information theoretic bounds on human knowledge of language

Frank Mollica & Steven T. Piantadosi

Abstract

We introduce theory-neutral estimates of the amount of information learners know about how language works. If children take years to extract a few bits of information, their learning mechanisms might be highly resource-limited or resource-intensive, as has been proposed in many flavors of nativism. On the other hand, if children extract large amounts of information in relatively short amounts of time, their learning mechanisms will need to be quite sophisticated. 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, corresponding to learners who extract 1000-2000 bits *per day* through the first 18 years of their life, suggesting the use of sophisticated learning mechanisms. Interestingly, the vast majority of stored information is lexical semantics.

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. Discovery of a lower bound on children's learning capacity would inform the debate between nature and nurture in that it would point to whether the most promising research direction would consider learners who operate on small parameter spaces (e.g., Wexler & Culicover, 1980) or large unrestricted spaces of hypotheses (e.g., Chater & Vitányi, 2007).

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 and measuring how far they were blown. His estimate—computed in his head, on the spot—was correct 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. These 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[1]. 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. This informs language learning theories: if the number of bits per day is low (a few bits per day) then language learning may best be characterized with extremely resource- and inferentially-limited systems, most consistent with the constraints of nativist theories. If the number is high (dozens, hundreds, or thousands of bits per day), that would point to remarkably sophisticated learning systems capable of extracting and synthesizing large amounts of data.

Our study is also 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) (Von Neumann, 1958). More recently, bounds for information transfer in neural coding have been estimated using information theoretic techniques (Borst & Theunissen, 1999). Working from behavioral performance, Landauer (1986) 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, Landauer (1986) 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 more recently by Brady, Konkle, Alvarez, and Oliva (2008) in a large-scale empirical study, which estimated that human memory could encode $2^{13.8}$ unique items. Landauer (1986) 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. 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 Hunter (1988), with a response given by Landauer (1988).

Our focus here is on estimating *linguistic* knowledge across multiple levels structure and function: phonemes, wordforms, lexical semantics, word frequency and syntax. At each level, there is a large space of logically possible linguistic units (e.g., acoustic cue values, syntactic parses). The challenge for learners is to discover and store which units are used in their language. Tools in information theory allow us to estimate the relevant quantities. First, we assume that before learning, children begin with a certain amount of uncertainty over the required representation, R, denoted $\mathbf{H}[R]$. Shannon entropy (Shannon, 1948) quantifies the number of bits that must be received on average to remove uncertainty about what R is the true one,

$$\mathbf{H}[R] = -\sum_{r\in R} P(r)\log P(r), \tag{1}$$

where r is each possible representation. For instance, R might be the space of grammars, r might be a particular grammar, and $P(r)$ is the probability, before learning begins that r will be the correct one (e.g. $P(r) = 1/|R|$ corresponds to equal expectations for all grammars). Information theory is powerful in large part because it applies regardless of what R is, and in some cases $\mathbf{H}[R]$ can be estimated without knowing the details of R itself.

After observing some data D, learners will have a new amount of uncertainty (perhaps zero) over R, denoted $\mathbf{H}[R \mid D]$,

$$\mathbf{H}[R \mid D] = -\sum_{r\in R} P(r \mid D)\log P(r \mid D). \tag{2}$$

It is a standard theorem in information theory (see Cover & Thomas, 2012) that $\mathbf{H}[R \mid D] \leq \mathbf{H}[R]$, meaning that providing data D must not increase the amount of information required to convey which R is the right one. We can formalize the amount of information that D provides about R, denoted $\mathbf{I}[R; D]$ as the difference between $\mathbf{H}[R]$ and $\mathbf{H}[R \mid D]$,

$$\mathbf{I}[R; D] = \mathbf{H}[R] - \mathbf{H}[R \mid D]. \tag{3}$$

This quantity is called "mutual information" and it gives the amount of information that D (e.g. data from learning) provides about a representation R. 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 $\mathbf{H}[R \mid D]$. The resulting value will tell us the number of bits of information that the learning data D has provided. In the case of continuous distributions, (1) through (3) have continuous analogs where the sums turn into integrals,

$$\mathbf{H}[R] = -\int P(r)\log P(r)\ dr \quad \text{and} \quad \mathbf{H}[R \mid D] = -\int P(r \mid D)\log P(r \mid D)\ dr, \tag{4}$$

with still $\mathbf{I}[R; D] = \mathbf{H}[R] - \mathbf{H}[R \mid D]$. Together, (1)-(4) will be used to quantify information in each domain of language.

2 Results

We will build up our estimates separately for each linguistic domain. The results of each section are summarized in Table 1. Table 2 summarizes the key assumptions behind each of our estimations.

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	200, 000	12, 000, 000	40, 000, 000
2.4	Word Frequency	40, 000	80, 000	120, 000
2.5	Syntax	134	697	1394
	Total	440, 512	12, 481, 447	40, 762, 849
	Total per day	67	1, 900	6204

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 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

$$\mathbf{H}[R] = -\int P(r) \log P(r)\, dr = \log(B - A), \tag{5}$$

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 $\mathbf{H}[R \mid D]$ as

$$\mathbf{H}[R \mid D] = \int N(x \mid \mu, \sigma) \log N(x \mid \mu, \sigma)\, dx = \frac{1}{2}\log(2\pi e \sigma^2) \tag{6}$$

where N is a normal probability density function, and μ and σ are the participants' inferred mean and standard deviation. Thus, the final entropy is just a simple function of the standard deviation of the representation, with lower standard deviations giving lower entropy. To find the information learned, we must also use (4), yielding that the amount of information learned for a single acoustic dimension is[2]:

$$\mathbf{I}[R; D] = \mathbf{H}[R] - \mathbf{H}[R \mid D] = \log(B - A) - \frac{1}{2}\log(2\pi e \sigma^2), \tag{7}$$

To find σ for real humans, we use the values inferred by Kronrod, Coppess, and Feldman (2016, Table 7) to account for the perceptual magnet effect[3].

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 (Anglin, Miller, & Wakefield, 1993). 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 $12,000-17,000$ (Zechmeister, Chronis, Cull, D'Anna, & Healy, 1995; D'Anna, Zechmeister, & Hall, 1991). Lexical knowledge extends beyond words too. Jackendoff (1997) 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 $\mathbf{H}[R \mid D]$ 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.[45]

To estimate these surprisals, we used the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995), we computed the surprisal of each word under 1-phone, 2-phone, 3-phone and 4-phone models (see Manning & Schütze, 1999) 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.[6] 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 (Laurence & Margolis, 1999). 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 leaner 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 (6) above gives the 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

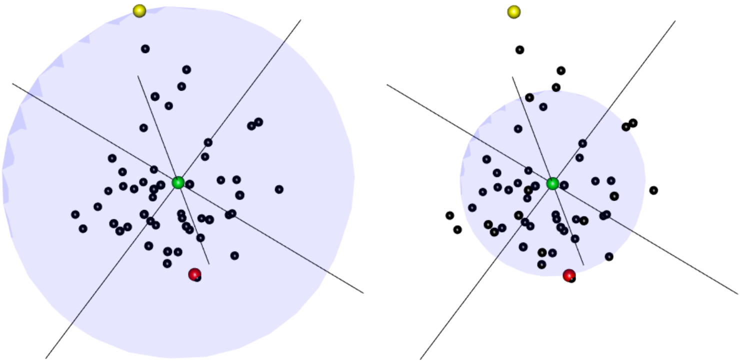

Figure 1: The shaded spheres represent uncertainty in semantic space centered around a word (in green). Left: The uncertainty is given with respect to the word's farthest connection in semantic space (in yellow), representing R. Right: The uncertainty is given with respect to the N^{th} nearest neighbor 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.

by:

$$\mathbf{H}[R] = \frac{N}{2}(1 + \log 2\pi + \log \sigma). \tag{8}$$

This means that if we go from a R standard deviation Gaussian to a r standard deviation one, the amount of information we have learned is the difference between these entropies,

$$\frac{N}{2}(1 + \log 2\pi + \log R) - \frac{N}{2}(1 + \log 2\pi + \log r) = \frac{N}{2} \log \frac{R}{r} \tag{9}$$

We estimate R and r in several different ways by looking at WordNet (Fellbaum, 1998) 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. As this shows, the range of likely values falls 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 Landauer and Dumais (1997), with numbers ranging from $100 - 500$ dimensions. Our low-bound will use 0.5 bits-per-dimension times only 10 dimensions, treating a word as essentially being spherical only on a subset of its semantic dimensions. This yields a lexicon total of $200,000$ bits. Our best guess will use 1 bit per dimension and 300 dimensions following Landauer and Dumais (1997) for $12,000,000$ bits. Our upper bound uses 2 bits-per-dimension and 500 dimensions for a total of $40,000,000$ bits.

2.4 Information about Word Frequency

Word frequencies are commonly studied in psychology as factors influencing language processing and acquisition (e.g., Forster & Chambers, 1973; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Plaut, McClelland, Seidenberg, & Patterson, 1996; Zorzi, Houghton, & Butterworth, 1998; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Murray & Forster, 2004) as well as for their peculiar Zipfian distribution (Piantadosi, 2014). However, relatively little work has examined the fidelity of people's representation

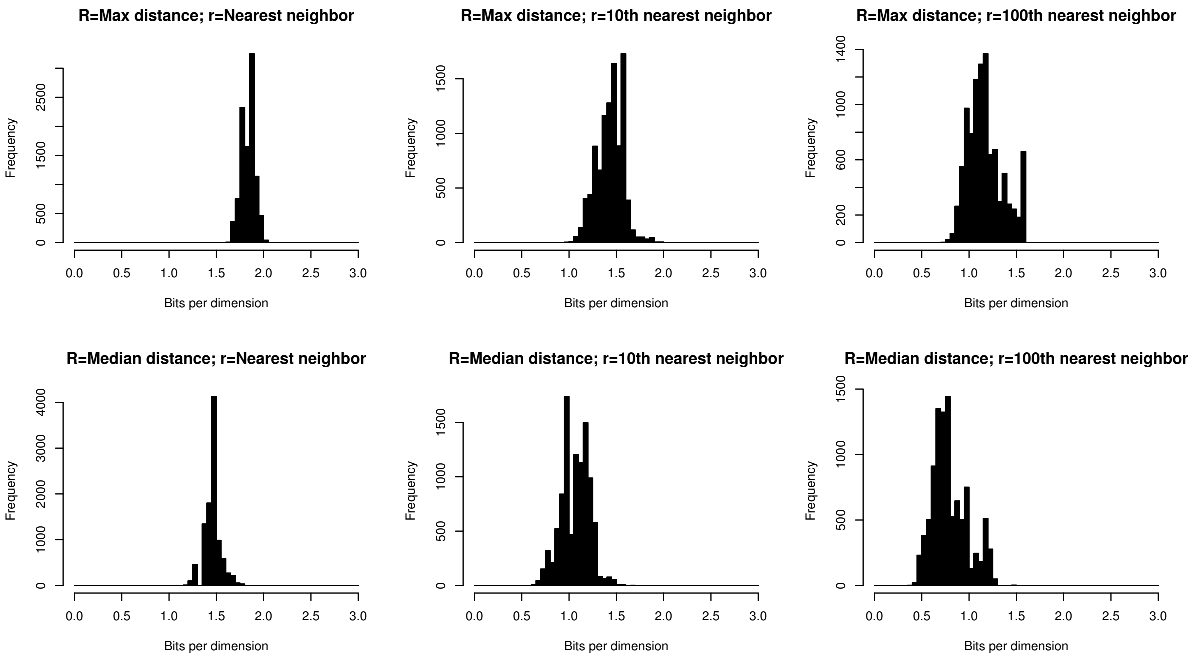

Figure 2: Histograms showing the number of bits-per-dimension $(\frac{1}{2}\log\frac{R}{r})$ for various estimates of R and r. These robustly show that $0.5 - 2.0$ bits are required to capture semantic distances.

of 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 $\log(40000!) \approx 500,000$ 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 frequency[7]. Words were sampled from the lexical database SUBTLEX (Brysbaert & New, 2009) in 20 bins of varying log frequency. We removed words below the bottom 30'th percentile (frequency count of 1) and words above the upper 99'th 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 (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

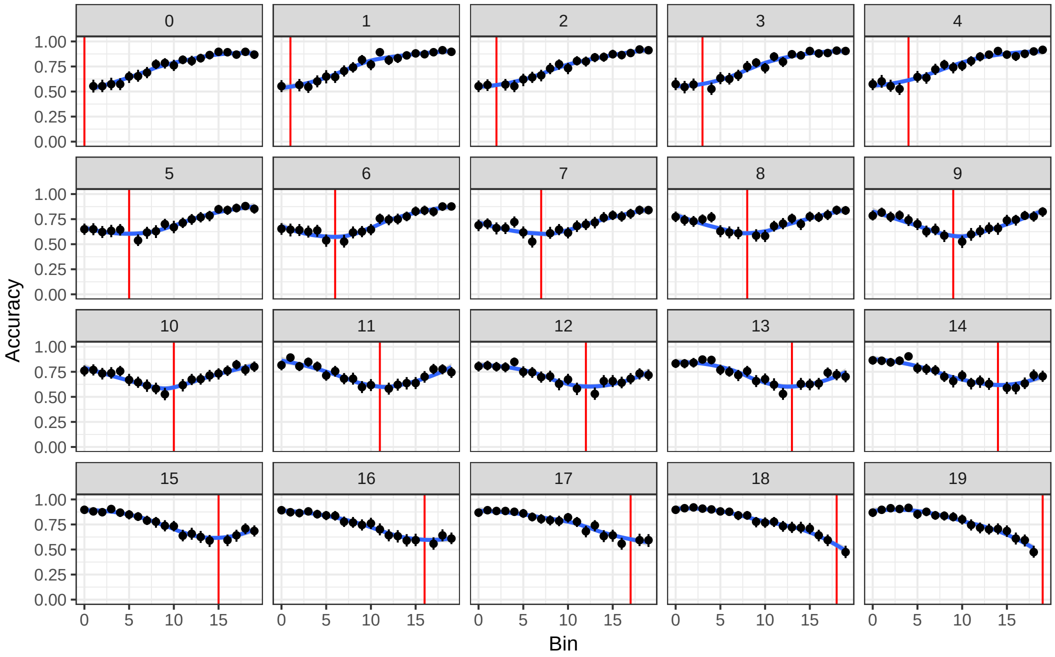

Figure 3: Accuracy in frequency discrimination accuracy as a function of log word frequency bin faceted by log reference word frequency bin. Vertical red lines denote within bin comparison. Line ranges reflect 95% bootstrapped confidence intervals.

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 (Wexler & Manzini, 1987; Kohl, 1999) to ones that consider alternative spaces of infinite models (e.g. Perfors, Tenenbaum, & Regier, 2011; Chater & Vitányi, 2007) or data-driven discovery from the set of all parse trees (Bod, Scha, Sima'an, et al., 2003). 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 $s_1, s_2, \ldots, s_n$ of sentences which share as little syntactic structure as possible between each s_i and s_j. For instance,

$$\text{Bill [met John]} \tag{10}$$

and

$$\text{[Jill's sister] cried} \tag{11}$$

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 theses "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

$$\mathbf{H}[R] = -\sum_{r \in R} \frac{1}{\#parses} \log \frac{1}{\#parses} = log(\#parses). \tag{12}$$

We assume the knowledge of an adult leaves zero uncertainty in general, yielding

$$\mathbf{H}[R \mid D] = 0 \tag{13}$$

so that

$$\mathbf{I}[R; D] = \mathbf{H}[R] - \mathbf{H}[R \mid D] = log(\#parses) \tag{14}$$

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

$$C_n = \frac{1}{n+1}\binom{2n}{n}. \tag{15}$$

As an example, to determine each of (10) and (11), 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

$$\sum_{i=1}^{n} \log C_{(l(s_i)-1)}. \tag{16}$$

Again, (16) assumes that there is no syntactic structure shared between the s_i—otherwise (16) 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 Sprouse and Almeida (2012). They present 111 sentences that are meant to span the range of interesting linguistic phenomena and were presented independently in Adger (2003). Our best estimate is therefore (16) taking s_i to be the lengths of these sentences. We take the lower-bound to be (16) 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 Sprouse and Almeida (2012) may not cover the majority of syntactic structures, particularly when compared to more exhaustive grammars like Huddleston, Pullum, et al. (2002). 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

By providing bounds on the amount of information language users must acquire to effectively use their language, we aim to better characterize the learning mechanisms supporting acquisition. Summing across our estimates for the amount of information language users store about phonemes, wordforms, 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 Landauer (1986)'s estimate for language knowledge. 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 400,000 bits, which implies that learners would remember around 67 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,

1001001111000100011001111110001101111101101100001100011110100100110.

Of course, 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 consistent with the view that language is supported by sophisticated learning mechanisms.

Importantly, our estimates vary on orders of magnitude across levels of representation. As these analyses show, the majority of information humans store about language is linked to words, specifically lexical semantics, as opposed to core systems of language knowledge, such as phonemes and syntax. These differences could suggest fundamental differences in the learning mechanism for specific language learning problems. 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 10 years, children would have to learn a single bit about syntax every five days on average. Despite this, the hypothesis space for learners in our best guess for syntax consists of $2^{697} \approx 10^{210}$ different hypotheses for how syntax works. In other words, learners still would need the ability to navigate an immense space of hypotheses, 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.

Of course 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 2). 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.

References

Adger, D. (2003). *Core syntax: A minimalist approach* (Vol. 33). Oxford University Press Oxford.

Anglin, J. M., Miller, G. A., & Wakefield, P. C. (1993). Vocabulary development: A morphological analysis. *Monographs of the society for research in child development*, i–186.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). Celex2 ldc96l14. *Web Download. Philadelphia: Linguistic Data Consortium.*

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283.

Bod, R., Scha, R., Sima'an, K., et al. (2003). *Data-oriented parsing*. Citeseer.

Borst, A., & Theunissen, F. E. (1999). Information theory and neural coding. *Nature neuroscience*, *2*(11), 947–957.

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*, *105*(38), 14325–14329.

Brysbaert, M., & New, B. (2009). 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.

Chater, N., & Vitányi, P. (2007). 'ideal learning'of natural language: Positive results about learning from positive evidence. *Journal of Mathematical psychology*, *51*(3), 135–163.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, *108*(1), 204.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory.* John Wiley & Sons.

D'Anna, C. A., Zechmeister, E. B., & Hall, J. W. (1991). Toward a meaningful definition of vocabulary size. *Journal of Literacy Research*, *23*(1), 109–122.

Fellbaum, C. (1998). *Wordnet.* Wiley Online Library.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of verbal learning and verbal behavior*, *12*(6), 627–635.

Huddleston, R., Pullum, G. K., et al. (2002). The cambridge grammar of english. *Language. Cambridge: Cambridge University Press*, 1–23.

Hunter, L. (1988). Estimating human cognitive capacities: A response to Landauer. *Cognitive Science*, *12*(2), 287–291.

Jackendoff, R. (1997). *The architecture of the language faculty* (No. 28). MIT Press.

Kohl, K. T. (1999). *An analysis of finite parameter learning in linguistic spaces* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, *23*(6), 1681–1712.

Landauer, T. K. (1986). How much do people remember? some estimates of the quantity of learned information in long-term memory. *Cognitive Science*, *10*(4), 477–493.

Landauer, T. K. (1988). An estimate of how much people remember, not of underlying cognitive capacities. *Cognitive Science*, *12*(2), 293–297.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. *Concepts: core readings*, 3–81.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). MIT Press.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, *111*(3), 721.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(5), 1112–1130.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological review*, *103*(1), 56.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 623–656.

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics*, *48*(03), 609–652.

Von Neumann, J. (1958). *Tbe computer and the brain.* Yale University Press.

Wexler, K., & Culicover, P. (1980). Formal principles of language acquisition.

Wexler, K., & Manzini, M. R. (1987). Parameters and learnability in binding theory. In *Parameter setting* (pp. 41–76). Springer.

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A., & Healy, N. A. (1995). Growth of a functionally important lexicon. *Journal of Literacy Research*, *27*(2), 201–212.

Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two routes or one in reading aloud? a connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(4), 1131.

Notes

[1] Hint: you can use your knowledge of how many miles car tires last for and how much thickness they lose over their lifetime.

[2] For vowels, we extend these distributions to their multidimensional counterparts as formant space is two dimensional.

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

4 Acknowledgements

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

[4] In 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.

[5] Analogously, 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.

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

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

Table 2: The assumptions we make in our estimates.

Section	Domain	Assumptions
2.1	Phonemes	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 Kronrod et al. (2016).
2.2	Phonemic Wordforms	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	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	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	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 Sprouse and Almeida (2012) are a good approximation/coverage of the essentially independent syntactic components of English grammar.