Language Learning as Codebreaking: The Key Roles of Redundancy and Locality

Richard Futrell
University of California, Irvine
rfutrell@uci.edu

Abstract

Understanding the inherent properties that render a language learnable remains a fundamental question in cognitive science and linguistics. I propose to analyze language learning as a codebreaking task, wherein the learner recovers the underlying grammar (the cryptographic key) from observed linguistic input (intercepted ciphertext). I develop a standard information-theoretic analysis of this codebreaking problem, but with a twist: in cryptography, one wants to make a code unbreakable, but in language, one wants the language to be learnable. The analysis yields three main findings: (1) Semantic redundancy—predictability of meanings given context—is necessary for language learning; (2) When learners have limited memory for sequential information, this redundancy must be local within linguistic strings; and (3) certain simple kinds of compositional languages naturally embody this kind of local semantic redundancy, enhancing their learnability. The framework shows how distributional statistics enable the learning of form–meaning mappings even when learners only observe forms.

1 Introduction

Theoretical models of language learning often focus on the knowledge that a human brings to the task, in the form of formal restrictions on possible grammars (Chomsky, 1965), simplicity biases (Hsu and Chater, 2010; Hsu et al., 2013), or Bayesian priors (Griffiths and Kalish, 2007; Pearl, 2023). Here I instead ask what properties of language make it learnable regardless of prior knowledge, based on a cryptanalytic approach: I consider the language learner to be a codebreaker attempting to infer a cryptographic key (the grammar of a language, which I take to include the lexicon) based on intercepted encrypted ciphertexts (linguistic input). I adapt the classic information-theoretic treatment of this codebreaking problem (Shannon, 1949) with a twist: whereas in cryptography one is interested in designing codes where the key is hard to break, here I treat language as a code that wants to be broken. The parallel language learning and codebreaking is illustrated in Figure 1.

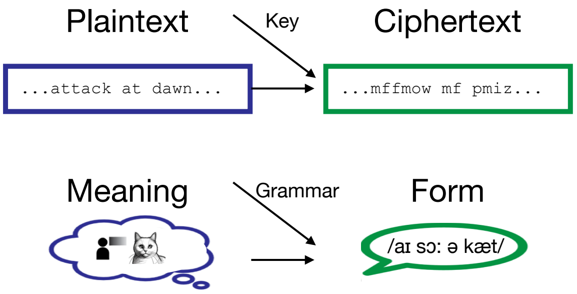

Figure 1: Parallel between language and cryptography. In cryptography (top row), a plaintext (a string) is encrypted using a secret key to form a ciphertext (another string). An attacker may determine the secret key by observing many ciphertexts; the system is designed to make this codebreaking task difficult. In language (bottom row), a meaning (in an arbitrary representational format) is expressed as a form (a string) using an unknown grammar. A learner may determine the grammar by observing forms; if the language is to be learnable, it should be structured so that this codebreaking task is easy.

I present three main results:

- Language learning crucially depends on *semantic redundancy* of the input.
- Given that learners have limited memory for sequences, this redundancy must be *local* within strings.
- Certain simple kinds of *compositional* languages exhibit exactly this kind of local redundancy and are more learnable as a result.

Furthermore, the cryptanalytic approach clarifies when and how semantics can be learned from distributional statistics (Harris, 1954; Mikolov et al., 2013; Merrill et al., 2021).

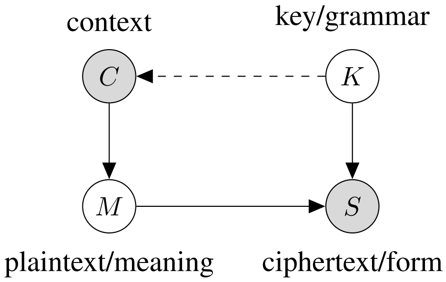

Figure 2: Probabilistic graphical model representation of the learning problem. Forms are a function of a key/grammar K and a meaning M. The learner observes context C and form S and tries to infer the key/grammar K. The learner never observes underlying meanings M. For extralinguistic context, there is no dependency of C on K. For intralinguistic context, there is such a dependency.

2 Language learning as codebreaking

Idealizing, let a **language** L_k be an injective mapping from **plaintexts/meanings** $\mathcal{M}$ to **ciphertexts/forms** which are strings drawn from a finite alphabet, parameterized by a **key/grammar** k, with each key corresponding to a unique possible mapping.[1] Let M be a random variable over meanings, K be a random variable over keys, and $S = L_k(M)$ be a random variable over forms derived by applying some language to meanings M. The context C may be extralinguistic (for example, the sensory context of a caretaker pointing to a ball before saying "ball") or intralinguistic (for example, the words "that red" appearing before "ball"). The structure of the problem is schematized in a probabilistic graphical model in Figure 2.

The main quantity of interest for the codebreaking problem is the **leakage rate**, the amount of information that each ciphertext sample S provides about the key K. In cryptography one wants to minimize the leakage rate, but when thinking about language learnability we will be thinking about how to maximize it. Leakage rate is formally the mutual information between ciphertexts and keys given context:[2]

$$L = I[S : K \mid C]. \quad (1)$$

Each intercepted ciphertext S leaks some information about the key. The number of bits of leaked information needed to break the code is (on average) the entropy over keys $H[K]$. Leakage rate tells us how quickly the code can be broken, that is, how much ciphertext the learner must intercept before they can learn the language / determine the key, a quantity called **unicity distance** (Shannon, 1949, p. 693).

Given this analysis, there are two ways to make a language learnable.[3] The first is to set up learners to have a restricted distribution over possible grammars, thus lowering $H[K]$, the amount of leaked bits that must be gathered to break the key. The second is to increase the leakage rate, that is, to speak a language where the average form is highly informative about the key, regardless of what the prior distribution on keys looks like. I will focus on this latter aspect of language learnability.

3 Semantic redundancy

The first result is that languages are learnable to the extent that meanings are more predictable than forms. I formalize this using the notion of **semantic redundancy**, the predictability of meanings given context. I operationalize semantic redundancy using the conditional entropy of meaning given context $H[M \mid C]$, which represents the uncertainty about meaning given context: lower conditional entropy means more semantic redundancy. We will see that a language is more learnable when this quantity is small, corresponding to high semantic redundancy. Semantic redundancy may be contrasted with **formal redundancy**, the extent to which a form is predictable given context, that is the extent to which the entropy on forms $H[S \mid C]$ is not maximal.

3.1 Derivation: The importance of semantic redundancy

The first result is that there is leakage when there is more uncertainty about form than about meaning:

Proposition 1. *For extralinguistic context C, the leakage rate L is equal to formal minus semantic entropy:*

$$L = H[S \mid C] - H[M \mid C]. \quad (2)$$

[1] In cryptography the plaintext is usually also a string, but this is not necessary for the information-theoretic analysis of codebreaking. In fact, the theory does not depend on any assumptions about the nature of the set of meanings $\mathcal{M}$.

[2] I assume familiarity with the information theory concepts of entropy and mutual information. See Cover and Thomas (2006, Ch. 2) for an introduction and reference.

[3] A reviewer suggests that *iconicity* also makes a language more learnable, for example if every word is represented by an onomatopoeic form. I believe this kind of iconicity is best thought of as a (soft) restriction on the prior over keys, such that languages containing certain iconic mappings have high prior probability.

Proof. Starting with the definition of leakage and applying standard information-theoretic identities (Cover and Thomas, 2006, Ch. 2), we get

$$L = I[S : K \mid C] \tag{3}$$
$$= H[S \mid C] - H[S \mid C, K] \tag{4}$$
$$= H[S \mid C] - I[S : M \mid C, K] - H[S \mid C, K, M]. \tag{5}$$

The last term is zero because $S = L_k(M)$ is a deterministic function given knowledge of the key k, and also we have $I[S : M \mid C, K] = H[M \mid C, K]$ because languages are injective. Finally, since keys K are independent of meanings M, we have $H[M \mid C, K] = H[M \mid C]$ and we arrive at (2). □

Remark 1. The argument depends on the fact that although the learner never has access to the true underlying meanings, they do have access to a *distribution* on meanings that they think are likely to be expressed.

Remark 2. This argument corresponds to the classic result that leakage rate is a function of redundancy per character of plaintext (Shannon, 1949, p. 689), but generalized. In the current setting, the analog to plaintexts is meanings M, but these are not necessarily expressible as strings. Shannon's result still holds, except instead of being phrased in terms of characters of plaintext, the analogous quantity is characters of ciphertext given the key (appearing in Eq. 4).

Remark 3. For intralinguistic context C, we can derive a similar form for leakage,

$$L = H[S \mid C] - H[M \mid C, K], \tag{6}$$

which differs only in that the semantic entropy is conditional on the key. This is because one can only 'unlock' the semantic redundancy in the intralinguistic context to the extent that one already knows the language. The interpretation of this quantity is largely the same as for extralinguistic context.

3.2 Why does redundancy enable learning?

There are two intuitions that elucidate why it is possible to learn a form–meaning mapping when there is a low entropy on meanings given contexts.

Intuition 1: Revealed meaning. Imagine a scenario where you know exactly the single meaning $m \in \mathcal{M}$ that will be conveyed, and receive a form $s \in \Sigma^*$. Then you can filter your distribution over languages to include the mapping $m \to s$, in addition to any other updates. This scenario is the extreme case where semantic entropy $H[M \mid C] = 0$. As $H[M \mid C]$ gets smaller, learning is more and more like this scenario: low entropy over meanings means that each utterance provides partial information about the full mapping. On the other hand, if the entropy over meanings is high, then no update or only a small update is possible.

Intuition 2: Dancing men. In *The Adventure of the Dancing Men* (Doyle, 1903), Sherlock Holmes encounters messages represented as strings of dancing men of different shapes. He deduces that this is a substitution cipher, where each English letter corresponds to a certain dancing man, and breaks the code by matching the dancing men to letters based on their statistical frequency of occurrence, the letter E being the most frequent letter. In general, a substitution cipher for English plaintexts can be broken by plotting a histogram of ciphertext letter frequencies against a histogram of English letter frequencies, and finding the mapping that makes the histograms match, an approach known as **frequency analysis**. This is possible because English letters are redundant, that is, the frequency distribution over English letters is relatively low entropy.

Similarly, given some string observations and some low-entropy distribution on meanings $H[M \mid C]$, corresponding to a highly skewed histogram, one can recover the key by matching the frequencies of strings in context with the probability distribution on meanings in those contexts. On the other hand, if the entropy of meanings $H[M \mid C]$ is high, then both the form frequencies and the meaning distribution will be close to flat, and so the histogram-matching approach will either not yield a unique solution, or will only work after intercepting a very large number of forms.

Distributional learning In distributional learning, one learns language entirely on the basis of frequency of occurrence and co-occurrence with context in the input. Distributional learning is a successful approach to modeling aspects of child language acquisition (Saffran et al., 1996) as well as developing computational representations of word meanings (Mikolov et al., 2013; Pennington et al., 2014). The result above clarifies why distributional learning works even when a learner never observes meanings directly (compare Bender and Koller, 2020): because intra- and extralinguistic contexts are informative about meaning, and thus can stand in as a proxy for meaning in an information-theoretic sense.

If language lacked semantic redundancy of this kind—that is, if $H[M \mid C]$ were maximal—then distributional learning would be impossible, as we would have $H[S \mid C] = H[M \mid C]$ and leakage $L = 0$. In fact, this corresponds to the notion of **perfect secrecy** in the cryptography setting (Shannon, 1949, §10), and optimal codes such as Huffman codes (Huffman, 1952), which minimize redundancy by design, also have minimal leakage. On the other hand, as long as the entropy of meanings $H[M \mid C]$ is not maximal (either due to context, or simply because the distribution on meanings is non-uniform), then we have nonzero leakage $L > 0$ and the learner will be able to get some information about the key.

3.3 Cognitive and linguistic significance

There are two linguistically significant interpretations of this result, depending on whether one thinks of the context C as extralinguistic or intralinguistic.

If C is extralinguistic, then the result shows the importance of the speaker's choice of which meanings to express in which contexts. Examples would include a child's caretaker pointing to a ball before saying "ball"—thus creating a context C which is highly predictive about the intended meaning M—or the caretaker choosing to name objects already present in the immediate environment, thus pedagogically choosing *meanings* M to fit the context C. Cognitively, the result requires that the child is able to infer communicative intent from context, at least to some extent, and more generally has some sense of what meanings are more or less likely. Learning is possible when meaning is low-entropy for the learner.

If C is intralinguistic, then the result shows the importance of the language itself being semantically redundant, as a function of both its grammatical structure and usage choices of the speaker. An utterance such as "My favorite vegetable is . . . " provides semantic redundancy by predicting certain semantic features of the following word (provided one has already worked out the meaning of "vegatable"). Languages with grammatical cues to semantic features, such as Bantu languages with rich noun class systems, provide similar information through grammatical means. Intralinguistic semantic redundancy corresponds to the familiar experience of being able to guess the meaning of an unknown word in context, for example when reading.

3.4 The role of formal redundancy

An interesting wrinkle is that *formal* redundancy is not helpful for learning in this highly idealized setting: leakage is upper bounded by the formal entropy $H[S \mid C]$. This means that, when the *form* S of some linguistic input is highly predictable from context, this *reduces* the amount of information that the input provides to a learner.

The role of formal redundancy and its relationship with semantic redundancy must be interpreted carefully. Formal redundancy does not simply mean that a form is predictable, it means that a form is predictable *on average across the learner's key distribution*. Effectively, when the learner has narrowed down the keys to some subset, and a form is totally predictable under all those keys, then there is formal redundancy without semantic redundancy, because observing the form is totally unsurprising.

Formal redundancy without semantic redundancy can arise from, for example, phonotactic constraints. For example, suppose that a language has phonotactics where every front vowel is followed by only front vowels, that is, it has vowel harmony; and suppose that a learner is aware of the concept of vowel harmony and has narrowed their space of possible languages/keys only to those that respect vowel harmony. Then when a front vowel occurs in the context of a front vowel, it is formally redundant: it is uninformative about *anything*, including the meaning.

4 Locality: Learning with noise

The argument above establishes that a learnable language must have semantic redundancy, but tells us nothing about the structure of that redundancy. Next I consider learners whose memory or attention for sequences is noisy, such that their observations effectively consist of contiguous substrings rather than full strings. Such noisy memory is characteristic of human children (Cowan et al., 1999; Gathercole et al., 2004; Luna et al., 2004). In this setting, I find that languages are more learnable when their intralinguistic redundancy is *local*, that is, when the meaning of a character or word is predictable given nearby characters or words.

4.1 Derivation: Effect of noise on learning

I now assume that with probability e, the context C is unavailable to the learner, with $L(e)$ being the leakage rate as a function of the context erasure rate e. The idea is that a learner with limited memory or attention might find themselves processing part of a string without knowledge of its context.

In order to understand how the leakage changes as a function of noise rate e, one can calculate the derivative of $L(e)$ with respect to e:

Proposition 2. *For extralinguistic context C, the derivative of leakage with respect to context erasure rate e is equal to the formal minus semantic mutual information:*

$$\frac{\partial}{\partial e}L(e) = I[S : C] - I[M : C]. \tag{7}$$

Proof. Let $\tilde{C}$ represent the random variable over noisy context, equal either to a true context or to a special erasure symbol E not in the support of C. The leakage as a function of erasure rate $L(e)$ comes out to

$$\begin{aligned} L(e) &= H[S \mid \tilde{C}] - H[S \mid \tilde{C}, K] && (8)\\ &= H[S \mid C] - H[S \mid C, K] && (9)\\ &\quad + eI[S : C] - eI[S : C \mid K] \\ &= H[S \mid C] - H[M \mid C] && (10)\\ &\quad + eI[S : C] - eI[M : C]. \end{aligned}$$

The derivative of (10) with respect to e is (7). □

Remark 4. The analogous result for intralinguistic context is

$$\frac{\partial}{\partial e}L = I[S : C] - I[M : C \mid K], \tag{11}$$

paralleling the intralinguistic version of Prop. 1.

The result means that as a context becomes more likely to be unavailable to the learner, the learnability of the language goes up in proportion to the formal redundancy contributed by that context, and down in proportion to the semantic redundancy contributed by that context. Intuitively, if the learner has no access to context, then the semantic redundancy contributed by context cannot help. In terms of language learnability, the upshot is that languages should be configured so that helpful semantically redundant context is likely to be available in practice: that is, somewhere in the string where it is not likely to be erased.

4.2 Locality from noise

Consider now a scenario where a learner takes in a string incrementally and, at each position, has some probability of randomly forgetting (or otherwise ignoring) the string prefix up to that point. This represents a learner who either has noisy memory for the sequence context, or who has had a lapse of attention and is starting to process a string somewhere in the middle. Then the learner effectively has perceptual *intake* (in the sense of Pearl, 2023) consisting of contiguous substrings, rather than full strings.

In that case, if there is some helpful semantic redundancy between two nonlocal parts of a string, then this redundancy is unlikely to help the learner, since the learner is unlikely to get a large enough substring to encompass all parts. On the other hand, semantic redundancy between local parts of the string is more likely to be available. The upshot is that for a language to be learnable under these circumstances, it must have information locality (Futrell and Hahn, 2022): any helpful semantic redundancy should be expressed in *local* parts of a form, so that a learner with noisy memory or attention who is only receiving contiguous substrings as input is able to detect that redundancy and learn from it.

The idea of local semantic redundancy is related to the concept of **diffusion** from cryptanalysis (Shannon, 1949, pp. 708–709). Diffusion is a desirable property for cryptographic ciphers, where the redundancy in the plaintext is dissipated into long-range correlations involving many parts of the ciphertext, so that a codebreaker must intercept and analyze a very large quantity of contiguous ciphertext in order to detect the redundancy and exploit it. For learnability, human languages should do the opposite of diffusion: they should be set up so that semantic redundancy is detectable without considering large amounts of context.

5 Simulations

The considerations above suggest that for languages to be learnable, (1) languages must have semantic redundancy, and (2) if there is noisy memory for sequence context, languages should configure strings so that semantically redundant parts are local. Here I demonstrate this result by simulating learning of some very simple languages which differ in their levels of redundancy, in the locality of that redundancy, and in the level of noise under

Meaning →	ⒽⒽⒽ	ⒽⒽⓉ	ⒽⓉⒽ	ⒽⓉⓉ	ⓉⒽⒽ	ⓉⒽⓉ	ⓉⓉⒽ	ⓉⓉⓉ
Compositional 1	aaa	aab	aba	abb	baa	bab	bba	bbb
Compositional 2	bbb	abb	bab	aab	bba	aba	baa	aaa
Holistic 1	aab	bbb	bba	aba	baa	bab	aaa	abb
Holistic 2	abb	bbb	bab	baa	aab	aba	aaa	bba

Table 1: Example languages for the coinflip world, used in simulations. Possible meanings (coinflip outcomes) are on the columns. In the 'compositional' languages, each character corresponds to an individual coin, as indicated by color. In the holistic languages, there is no such correspondence.

which learning takes place. In line with the formal results, I find that semantic redundancy facilitates learning, and that in the presence of noise this redundancy must be local. Furthermore, I show how local redundancy obtains when languages are compositional in the sense that individual characters or local groups of characters (that is, words or morphemes) correspond to independent components of meaning.

5.1 Setup

I simulate ideal learners who start with an initial uniform distribution over keys/languages, observe (noisy) sample forms one at a time, and update their distribution on keys using Bayes' rule (Bayes, 1763).

Source As the probability distribution over meanings, I consider a very simple world consisting of two or three weighted coinflips, for a total of $2^2 = 4$ or $2^3 = 8$ possible outcomes/meanings. The first coin has weight b for heads, where I vary the weight b in order to vary the entropy of meanings $H[M]$—more biased coins yield lower-entropy distributions which should facilitate learning. The second and third coins have weights $b+0.1$ and $b+0.2$ respectively. If the coins did not have different weights, then the language would be unidentifiable for the learner, because the learner would never be able to identify which characters in a form correspond to which coins.

Languages I first consider languages where forms consist of binary strings of length 3, which are either compositional or not, in the sense that individual characters in the forms may or may not correspond to the underlying coinflips. These languages are categorized with examples in Table 1. I also consider redundant languages where forms consist of binary strings of length 4 and meanings consist of two coinflips. These languages are based on the Compositional 1 language in Table 1, and are either locally redundant (for example, a meaning ⒽⓉ is encoded as aabb) or nonlocally redundant (for example, the same meaning is encoded as abab). In all conditions, the learner's set of possible languages/keys is the set of all possible injective mappings from meanings to binary strings of the appropriate length.

Learning and noise In each step of learning, a learner observes a single (noisy) sample of a form, and updates their probability distribution on meanings exactly following Bayes' rule. Noisy observations are generated by sampling a form, splitting it into contiguous substrings, and uniformly choosing one of those substrings. The splitting is done by flipping a coin with probability e at each character of the string; if the outcome is heads, the string is split at that point. I vary the parameter e in experiments. The condition $e = 0$ corresponds to no noise. The condition $e = 1$ yields to a learner who only ever sees a single character of input based on a sampled string, corresponding to maximally noisy memory for intralinguistic context.

Evaluation I evaluate learning in terms of **key entropy**, the posterior entropy over keys given data observed so far at each timestep. Lower key entropy indicates the learner has less uncertainty about the language. The main feature of interest is the rate at which this entropy decreases.

I would like to emphasize that for all conditions in these simulations, the key entropy will eventually approach zero with enough observations: that is, learning is ultimately possible for all the languages considered here. They will differ, however, in their rates of learning.

5.2 Analysis of languages

The compositional languages in Table 1 have semantic redundancy local to each individual character. This is because the meaning of each character corresponds to one coinflip, and thus the semantic entropy for a single character is bounded: it cannot exceed the entropy of its corresponding single

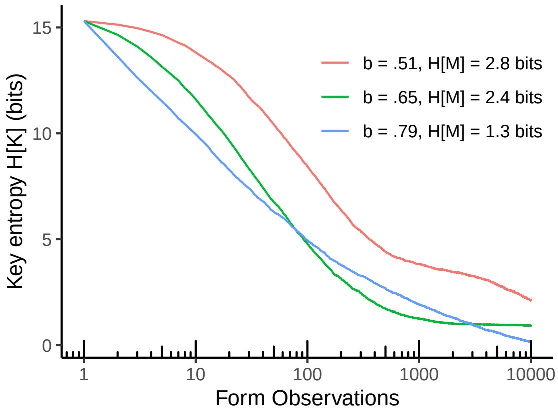

Figure 3: Learning curves (average over 10,000 runs) for different levels of semantic entropy, with no noise. Curves show key entropy $H[K]$ as a function of the number of forms observed (similar to Shannon, 1949, Fig. 6). Key entropy decreases more rapidly when semantic entropy is low. Curves are the same for all languages in Table 1.

coinflip. This redundancy is local in the sense that it does not depend on context and cannot be destroyed by erasure noise. On the other hand, in the holistic languages, each character corresponds to a *mixture* of different coins, which will generally have a higher entropy (thus less semantic redundancy) than the distribution of a single coin. Furthermore, there will be nonlocal correlations among the characters within the string, representing nonlocal semantic redundancy which is in danger of being missed due to noise. This observation is in line with the idea that noncompositional languages very generally create undesirable long-term correlations within forms (Futrell and Hahn, 2024).

The locally redundant variant of the compositional language extends this idea so that redundancy is local to a *pair* of adjacent characters. The helpful semantic redundancy in this adjacent pair is unlikely to be disrupted by noise, and thus learning curves are favorable. On the other hand, in the nonlocally redundant language, the redundancy is nonlocal, highly likely to be disrupted by noise, and so the learning curves are less favorable.

5.3 Results

Learning curves without noise ($e = 0$) by semantic entropy are shown in Figure 3, which demonstrates that learning is indeed faster when semantic entropy is lower. The language used for this simulation is Compositional 1 from Table 1, but this does not matter: in this setting, all injective languages will produce equivalent curves when there is no noise.

Learning curves under varying levels of noise are shown in Figure 4. Here we find that the compositional languages yield faster learning, as expected, because their semantic redundancy is local and not likely to be disrupted by noise. The difference between compositional and holistic languages gets bigger as the noise rate increases. Learning curves for the explicitly redundant languages are shown in Figure 5. Languages with local redundancy are faster to learn, while languages with nonlocal redundancy are slower.

6 Discussion and Related Work

I emphasize that I have considered learners who never directly observe meaning, and who have no *prior* bias towards any language over another; nor is any language ‘simpler’ than any other for the learners. The fact that certain languages are learned more rapidly is rather a function of their semantic redundancy and information locality, which enables learning in the presence of noisy memory or attention for sequences, in a way that is independent of the learner’s prior distribution over languages.

Distributional learning This work provides a theoretical understanding of when it is possible to learn a form–meaning mapping from observations of form alone, and thus justifies distributional approaches to semantics and language learning (Harris, 1954; Erk, 2010), both in the context of language technologies (Mikolov et al., 2013), and as a strategy for child learners (Saffran et al., 1996; Erickson and Thiessen, 2015). The results are consistent with Merrill et al.’s (2024) finding that corpus statistics encode entailment relations under the assumption that speakers are redundant, and I believe the notion of local semantic redundancy is likely related to Merrill et al.’s (2021) notion of semantic transparency, which is a precondition for distributional learning of semantics.

Language acquisition The model shows how language can be acquired when context provides partial information about meanings, and thus it provides a generalized idealized version of the cross-situational learning model of lexicon acquisition (Siskind, 1996; Hendrickson and Perfors, 2019), in which a child encounters a word across multiple contexts until they can identify the word with a single meaning by a process of elimination. The re-

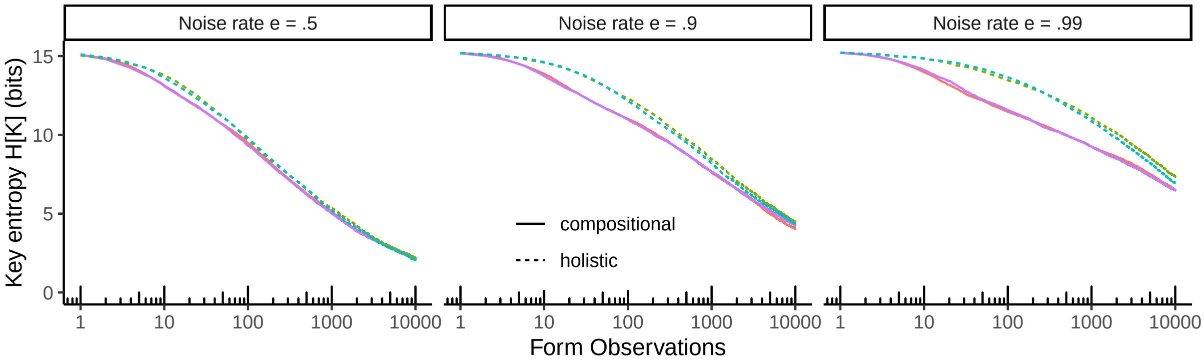

Figure 4: Learning curves for different levels of noise e, for a source with a fixed $b = .75$ (average over 10,000 runs). Curves show key entropy $H[K]$ as a function of the number of forms observed. Key entropy decreases more rapidly for the compositional languages, where semantic redundancy is local. It increases more slowly for the holistic languages where semantic redundancy is spread out among characters of the form. The difference between compositional and holistic languages is heightened for increased noise rates.

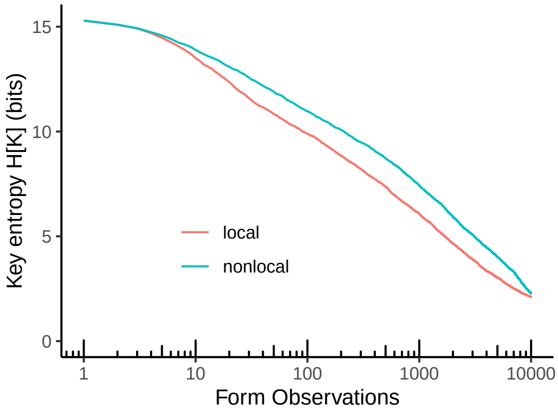

Figure 5: Learning curves for locally redundant and nonlocally redundant languages (see text) under noise at rate $e = .9$, for coinflip heads probability $b = .79$ (average over 1000 runs). Key entropy decreases more rapidly for the locally redundant languages.

sults about the importance of low semantic entropy are in line with the finding that children learn word meanings better given low-entropy input (Lavi-Rotbain and Arnon, 2019). The results on noise and locality show how cognitive constraints, such as maturational constraints on working memory, can imbue learners with a bias toward the kinds of structures found in language (Newport, 1990; Mita et al., 2025).

Unsupervised machine translation This work bears a notable similarity to models of how one can learn to translate between languages without seeing parallel texts (Cao et al., 2016), or how one might decode unknown communication systems such as those used by whales, where the nature of the meanings being expressed is unknown and possibly unknowable (Goldwasser et al., 2023). The current approach to language learning can be seen as inducing an unsupervised translation system from meanings (represented in some unknown mental form) to forms (represented as observable strings).

Language evolution Approaches to modeling language evolution by iterated learning have yielded the result that languages will generally reflect learners' prior distribution on languages (Griffiths and Kalish, 2007; Kirby et al., 2014). In contrast, I find a learning bias (toward locally redundant languages) as a function of the noisy nature of learners' intake, independent of the prior. This bias can be seen as arising from the learners' likelihood function rather than the prior, and it manifests in the *rate* of learning, not in its initial or asymptotic states. Under noise, locally redundant languages can be learned to a higher degree of confidence from fewer samples.

While humans may have innate prior knowledge of what grammars/keys are possible, the question remains of why that prior knowledge is what it is. For example, if humans' prior knowledge can be characterized by a constraint that languages must be compositional in a certain way, the question is why that constraint rather than another. The considerations above provide a potential explanation, by showing how learning biases can emerge independently of learners' priors. One could imagine a population of learners with flat priors, who end up with local compositional languages due to general memory limitations, as discussed in Section 4. Then over generations of evolutionary time, the population can evolve to incorporate these biases as innate prior knowledge.

7 Conclusion

I have presented a model of language learning based on ideas from cryptanalysis, in which a learner observes only forms and infers the underlying language, the mapping from hidden meanings to forms. Whereas in cryptanalysis one is concerned with making codes unbreakable, here I considered what properties of languages make them *breakable*. I found that languages with local semantic redundancy—the opposite of cryptographic diffusion, and corresponding to a kind of compositionality—are more learnable in this setting, even for learners without prior biases toward such languages. The model shows how learning is possible as long as the learner has some prior knowledge of their interlocutor's likely communicative intent.

The analytical and modeling approach taken here provides a useful new angle on language learning which can be applied to test hypotheses about how learning works, how properties of language affect learnability, and how the learner's hypothesis space on languages could be structured to enable rapid learning. More broadly, I believe that this cryptography-inspired analysis of language learning offers a fresh perspective and set of analytical tools that can be used to approach the language learning problem. Cryptanalysis is a well-developed and rich field of science and engineering. The analysis here shows that it may contain useful ideas for linguistics and language acquisition.

Acknowledgments

I thank Greg Scontras, Lisa Pearl, Connor Mayer, the UCI QuantLang Collective, and the three anonymous reviewers for helpful comments.

Code availability

Code to reproduce the simulations and figures is available at `http://github.com/langprocgroup/locallearning`.

References

Thomas Bayes. 1763. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S., communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53:370–418.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. 2016. A distribution-based model to learn bilingual word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1818–1827, Osaka, Japan.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ.

Nelson Cowan, Lara Nugent, Emily M. Elliott, Igor Ponomarev, and John Scott Saults. 1999. The role of attention in the development of short-term memory: Age differences in the verbal span of apprehension. *Child Development*, 70(5):1082–1097.

Arthur Conan Doyle. 1903. The adventure of the dancing men. *The Strand Magazine*, 26(156):603–617.

Lucy C. Erickson and Erik D. Thiessen. 2015. Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, 37:66–108.

Katrin Erk. 2010. What is word meaning, really? (And how can distributional models help us describe it?). In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 17–26, Uppsala, Sweden. Association for Computational Linguistics.

Richard Futrell and Michael Hahn. 2022. Information theory as a bridge between language function and language form. *Frontiers in Communication*, 7:657725.

Richard Futrell and Michael Hahn. 2024. Linguistic structure from a bottleneck on sequential information processing. *arXiv preprint arXiv:2405.12109*.

Susan E. Gathercole, Susan J. Pickering, Benjamin Ambridge, and Hannah Wearing. 2004. The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, 40(2):177–190.

Shafi Goldwasser, David Gruber, Adam Tauman Kalai, and Orr Paradise. 2023. A theory of unsupervised translation motivated by understanding animal communication. In *Advances in Neural Information Processing Systems*, volume 36, pages 37286–37320. Curran Associates, Inc.

Thomas L. Griffiths and Michael L. Kalish. 2007. Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31(3):441–480.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Andrew T. Hendrickson and Andrew Perfors. 2019. Cross-situational learning in a Zipfian environment. *Cognition*, 189:11–22.

Anne S. Hsu and Nick Chater. 2010. The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science*, 34(6):972–1016.

Anne S. Hsu, Nick Chater, and Paul Vitányi. 2013. Language learning from positive evidence, reconsidered: A simplicity-based approach. *Topics in Cognitive Science*, 5(1):35–55.

David A. Huffman. 1952. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101.

Simon Kirby, Thomas L. Griffiths, and Kenny Smith. 2014. Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28C:108–114.

Ori Lavi-Rotbain and Inbal Arnon. 2019. Children learn words better in low entropy. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 631–637.

Beatriz Luna, Krista E. Garver, Trinity A. Urban, Nicole A. Lazar, and John A. Sweeney. 2004. Maturation of cognitive processes from late childhood to adulthood. *Child Development*, 75(5):1357–1372.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

William Merrill, Zhaofeng Wu, Norihito Naka, Yoon Kim, and Tal Linzen. 2024. Can you learn semantics through next-word prediction? The case of entailment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2752–2773, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Masato Mita, Ryo Yoshida, and Yohei Oseki. 2025. Developmentally-plausible working memory shapes a critical period for language acquisition. *arXiv preprint arXiv:2502.04795*.

Elissa L. Newport. 1990. Maturational constraints on language learning. *Cognitive Science*, 14(1):11–28.

Lisa Pearl. 2023. Computational cognitive modeling for syntactic acquisition: Approaches that integrate information from multiple places. *Journal of Child Language*, 50(6):1353–1373.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Claude E. Shannon. 1949. Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4):656–715.

Jeffrey M. Siskind. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91.