Form is Function: Faithful Transmission of Information as a Pressure for Linguistic Structure

B005981
M.Sc. Evolution of Language and Cognition
The University of Edinburgh
2011

I have read and understood the University of Edinburgh guidelines on plagiarism and declare that this written dissertation is all my own work except where I indicate otherwise by proper use of quotes and references.

Acknowledgments

I cannot pretend to be the sole author of this work, whatever its worth; I have received constant guidance and support from my supervisors, Dr. Kenny Smith and Dr. Richard Blythe, to whom I am deeply indebted. I am particularly grateful to Kenny for his programming assistance – if I can write a coherent line of code, it is a testament to his tutelage. I also benefited from a long-distance programming tutor, Scott Wehrwein, who deserves a lifetime's supply of tea for his unfailing patience. Many, many thanks to Scott for responding to my inane queries with insight and enough sardonic humor to take the sting out of the umpteenth round of de-bugging.

I have been fortunate to enjoy the intellectually stimulating environment of the LEC and the support of my ELC coursemates, who were kind enough to tolerate the erstwhile mathematician in their midst. Such an active, enthusiastic academic community is a constant inspiration. My parents, who always nod and smile politely when I describe my progress, deserve great credit for encouraging me in my bizarre interdisciplinary studies, as does my brother, who has endured many conversations about probability, statistics, and the nature of mathematical proof.

Last but not least, I am indebted to the New Scotland Country Dance Society and the Dunedin Dancers for providing a welcome outlet for my non-academic energy and all my extra baked goods. Without NSCDS I wouldn't know an allemande from a petronella, and I offer them my heartfelt thanks for keeping me on my toes.

Philosophy is written in this great book – I mean the universe – which stands continually open to our gaze, but it cannot be understood unless one first studies the language and the characters in which it is written. It is written in the language of mathematics, and its characters are triangles, circles, and other geometrical figures, without which is it humanly impossible to understand a single word of it.

Galileo Galilei

Abstract

We take descriptions of language structure and evolution in terms of information theory as the basis of an investigation of the information-preserving capacity of linguistic features. We begin by examining factors affecting the performance of conventional error-correcting codes, which we compare to message construction algorithms designed to replicate basic linguistic features. We then subject these message construction algorithms to selective pressure maximizing transmission accuracy and examine the structure of the resulting systems. The emergence of structural patterns in the optimized systems points to the centrality of information preservation as a pressure on emerging linguistic systems.

Contents

Chapter 1

Introduction

Since the advent of writing, authors have encoded their messages to avoid distortion and misinterpretation. True codes, like the famous Caesar cipher, date back to the Romans, but implicit encoding strategies were present in the writing systems of ancient civilizations. The Egyptian hieroglyphic language, for example, manifested an early example of overt linguistic redundancy: the hieroglyphic lexicon contained several symbols for each syllable, all of which would be included in a transcription of that syllable so the reader could not fail to comprehend the message (Cherry 1951: 383-384). Even the peoples of antiquity designed their writing systems to facilitate comprehension under adverse circumstances.

The issue the ancient Egyptians combated through syllabic redundancy is no less salient today: how can we formulate a message to make sure it is correctly interpreted, no matter who receives it or what has happened to it since we wrote it? Claude Shannon formalized the study of communication under these conditions in his seminal 1948 article "A mathematical theory of communication."

Shannon's central problem was the following: channels through which messages are sent, like telephone wires or magnetic fields, often corrupt the data they carry. He sought to determine whether it was possible to rearrange the data in such a way as to preserve its original form under the "noise" the channel introduced. His conclusion, memorialized in the "noisy coding theorem," is a resounding yes. For a channel of fixed capacity, that is, a medium of information transfer through which a fixed number of bits of information can flow simultaneously, there is a set of codewords providing arbitrarily reliable communication for any rate of information transmission below the channel's capacity (Shannon 1948: 37). Codes make information transfer reliable.

1.1 Linguistics and Information Theory

Although Shannon's work is primarily relevant to computer science and mathematics, a small cadre of linguists, most notably Zellig Harris, also recognized its significance. Harris was the first to describe language in terms of the properties of information theory, implicitly hypothesizing the existence of correlations, if not outright equivalences, between the two systems.

Harris' mathematical reformulation of language, first propounded in his 1968 book *Mathematical Structures of Language*, draws heavily from Shannon's vast oeuvre. For Harris, the central properties of language derive from its semi-lattice structure, a structure Shannon had previously invoked for information (Shannon 1950a: 181). Harris takes what he calls the deviation from equiprobability – the fact that not all linguistic combinations occur with equal likelihood, and that some combinations never occur – as the point of departure for his description of language as a mathematically interdependent system. In English, for example, definite articles are unlikely to precede temporal or locative markers; we hear "the library" far more often than "the now," although there are phrases, like "in the here and now," where such combinations are accepted. An acceptable sentence featuring a phrase like "the yesterday" or "the somewhere" is even less likely, perhaps impossible. The linguistic importance of variably probable combinations like these prompts Harris to envision linguistic utterances as a probabilistic system in which "the presence of a word . . . depends not on particular other words but on the dependence properties of the other words" (Harris 1991: 58). Linguistic constraints are not strict rules but nested probabilistic relationships.

The interdependence of events is a central to Shannon's vision of information systems. For Shannon, the amount of "information" present in a system is a function of the probabilistic relationships obtaining among possible events, a concept he defines mathematically as *entropy* (Shannon 1948: 18). Shannon entropy, intuitively equivalent to uncertainty, is maximized when all events are equiprobable and minimized when one particular event is guaranteed to occur (op. cit: 19). Harris incorporates these upper and lower bounds in his delineation of the information capacity of linguistic structures. The likelihood of a combination is, in Harris' description, directly proportional to its informational value (Harris 1991: 54). In fact, the options available under a given constraint determine the informational capacity of each option. Each linguistic choice contributes to the amount of information conveyed by the utterance, and the amount of information added to the system is dependent on the laxity of the constraint (op. cit: 335). In essence, the more choice available in selecting a linguistic structure, the more information that structure carries. A wholly pre-determined structure, on the other hand, carries no information, just as in Shannon's mathematical formulation of entropy. By depicting language in terms of its informational properties, Harris highlights the striking resonances between linguistic and informational systems, implying a deeper relationship between the two domains.

Plotkin and Nowak (2000) take Harris' intuitive link between language and information a step further, postulating strict mathematical equivalences between the two systems. They begin by presenting a mathematical formulation of primitive communication in which the fitness of a communicative system is a function of several probabilistic sub-systems, as in Harris' redescription of language and Shannon's vision of information structure. The payoff of any communicative event is given by the probability that the speaker produces some signal j to refer to a meaning i, signal j is misconstrued as some signal k, and signal k is nonetheless interpreted as meaning i, resulting in a successful interaction (Plotkin and Nowak 2000: 149). In essence, a communicative system's fitness is a function of the number of ideas that can be conveyed successfully when subjected to diverse sources of noise, as in Shannon information theory. They go on to prove that this fitness function is mathematically equivalent to Shannon's definition of the fitness of a set of codewords (op. cit.: 153). Just as Harris successfully applied principles of Shannon entropy to a description of language, Plotkin and Nowak demonstrate that pressures affecting the fitness of information theoretic systems are equally applicable to linguistic systems.

1.2 A Causal Conundrum

Taking Harris' and Plotkin and Nowak's results together, language appears to be explicitly informational, structurally equivalent to Shannon's systems of codewords. The idea that natural language can be described using information theory is far from new; Shannon himself explored the analogy between natural language and codes, demonstrating in a 1950 paper that the entropy and redundancy of printed English can be approximated through higher-order Markov chains (Shannon 1950b: 194). As compelling as these comparisons are, they do not attempt to probe the causes of the resemblance, leaving us with an intriguing question: why?

Why do language and information exhibit such striking structural similarities? Is the similarity superficial, or is it a vestige of evolutionary pressures on the development of linguistic systems? Plotkin and Nowak's demonstration of the equivalence between communicative and coding fitness points to a deeper relationship, highlighting the sensitivity of both systems to common pressures. If, as Plotkin and Nowak assert, linguistic and informational systems share fitness properties, a system that is informationally optimal must also be linguistically optimal. Language, then, should replicate not only the structure but also the function of information theory: linguistic structure should conserve information under noise.

We contend, as a natural consequence of Plotkin and Nowak's result, that linguistic features should approximate the information-conserving properties of error-correcting codes, which reformulate messages in order to facilitate the identification and correction of errors in information transmission. Empirically, then, it is necessary to demonstrate that aspects of linguistic structure – fixed lexical or phonetic inventories, word-order conventions, and non-adjacency constraints, for example – enhance the preservation of messages exposed to noise. We support this theory by developing a series of increasingly complex language-like message construction algorithms. We asses these algorithms according to an information theoretic paradigm emphasizing transmission accuracy, then observe the effect of random variation in refining these systems. Finally, we evaluate the extent to which imposing basic probabilistic constraints induces higher levels of structure in linguistic output.

In designing our language-like message construction algorithms, we focus on two basic properties of linguistic communication: fixed segment inventories and statistical relationships among segments. We begin by discussing relevant literature pertaining to the importance of segment inventories and structural attributes that enhance their communicative efficacy.

Zellig Harris emphasizes the importance of pre-set linguistic conventions like lexical and phonemic inventories, arguing that reliance on a discrete rather than continuous system reduces the probability of error-compounding in message transmission (Harris 1991: 148). If an agent receives a continuous stream of data and attempts to imitate it, various aspects of the input, like prosody, are easily lost or altered, and these alterations are likely to be exaggerated when the data are imitated by another agent, and then another, and so on. If the agent receives discrete data, by contrast, he can use his knowledge of the set of possible units to reconstruct items he misperceives. For Harris, limiting utterances to pre-set segments makes the problem of error-compounding tractable and limits the impact of individual differences in articulation (op. cit.: 148).

According to Harris, any set of conventional segments should increase transmission accuracy, but computational research emphasizes the importance of maximally distinct segments. Liljencrants and Lindlbom (1972) were among the first to recognize the primacy of mutual distinctiveness. Hypothesizing that distinctiveness was a key pressure motivating the development of sound systems, they performed a computational simulation in which they evolved vowel systems of varying sizes. Maximizing the Euclidean distance between points in a three-dimensional space, each dimension representing one of the first three formants, allowed them to evaluate the resemblance between their optimized systems and those observed in modern languages (Liljencrants and Lindblom 1972: 842). Systems of three to twelve vowels were developed in this way and compared to modern sound systems; although there was not a complete correspondence between the predicted and naturally-occurring vowel systems, in the vast majority of cases the similarity between the two distributions was significant (op. cit: 854). Distinctiveness is thus a primary pressure on the formation of sound systems.

Liljencrants and Lindblom's emphasis on spatial distance between vowels does not, however, map directly to perceptual discriminability, as Zuidema and de Boer (2009) argue (127). Zuidema and de Boer modify Liljencrants and Lindblom's spatial model to incorporate not only spatial but also temporal distance (ibid). They refine the optimization heuristic to minimize the probability of confusing two signal trajectories and assume the existence of selective pressure on distinctiveness (ibid). Using this model, they demonstrate that optimizing perceptually discriminable trajectories results in combinatorial phonology, with trajectories re-using certain areas in the space (op. cit: 133). Zuidema and de Boer then prove their optimization model to be an evolutionary stable strategy leading to convergence on optimal phonetic repertoires (op. cit: 138). When phonemic systems evolve under conditions favoring maximal discriminability, combinatorial phonology emerges and represents an optimal ESS. De Boer has also shown that optimizing imitation in a population of simulated agents produces realistic sound systems in which vowel clusters are optimally distributed in the acoustic space (de Boer 2000: 33), indicating that interaction among agents produces similar sound systems.

Zuidema and Westermann (2003) present similar results on the importance of maximal distinctiveness as a characteristic of the optimal lexicon. They begin by formulating a measure of communicative success, C, based on the probability that a given agent will accurately convey a certain meaning to another agent (Zuidema and Westermann 2003: 389). For each of several conditions, they initialize a set of agents with random form-meaning mappings and allow them to interact; throughout the interactions, the agents refine their form-meaning mappings according to a hill-climbing algorithm retaining only alterations that improve overall communicative success (op. cit: 390). When inter-agent communication is subject to noise, the meaning-form mapping that emerges is composed of maximally distinct meaning-form clusters (op. cit: 391). Maximal distinctiveness is thus a key component not only of optimal sound systems but also of optimal lexica.

Maximally distinct linguistic units are clearly crucial to communication, but relationships among those units are equally important. As Shannon and Harris emphasize, probabilistic interdependencies lie at the heart of both information and language structure. The importance of probabilistic computations in language perception throws the centrality of this property into sharp relief and suggests another fruitful avenue of investigation.

Research in first language acquisition further underlines the importance of such statistical analysis in distinguishing permissible from impermissible words. This ability is present from an early age; having demonstrated that eight-month old infants employ probabilistic techniques to segment continuous speech streams (Saffran et al. 1996: 1927), Aslin et al. (1998) further specify the statistical computation eight-month-old infants employ, pinpointing transitional probability as the key calculation. After being exposed to a sound sequence of four trisyllabic nonsense words over the course of three minutes, the infants were presented with twelve trisyllabic sound sequences (Aslin et al. 1998: 322-323). Of these twelve sound sequences, six were words from the sample corpus and six were composed of the final syllable of one corpus word and the first two syllables of another corpus word (op. cit: 323). Each test sequence was coordinated with a visual stimulus that allowed the experimenters to time the infant's attention to the sequence (ibid). Looking times were significantly longer for part-words than words, a result that Aslin et al. attribute to the novelty of these items based on the low transitional probability with which such a combination would have appeared in the training sample (op. cit: 324). Probabilistic computations underpin linguistic perception.

In fact, it is exactly this type of reasoning that Shannon employs to define information sources and approximate natural language mathematically. In information theory, the source of a message is represented as a discrete-time Markov chain specifying probabilities of transition from one symbol to another. The transition matrix defining the Markov chain encodes statistical information comparable to that accumulated by the infants in Aslin et al.'s experiment; each row in the matrix represents a possible state in the system, and each entry in that row represents the conditional probability of a transition to the state represented by the corresponding column. Varying the definition of a "state" in the chain allowed Shannon to develop increasingly accurate mathematical representations of English (Shannon 1950b: 202), indicating that probabilistic analysis not only facilitates language perception but also replicates language's mathematical structure.

Substantial evidence from the literature supports our choice of message construction algorithms; fixed segment inventories and their properties, coupled with statistical relationships among segments, are essential to linguistic communication. We hope to demonstrate that these fundamental aspects of language not only facilitate communication but also preserve information in the presence of noise.

If, as we maintain, language exhibits properties of error-correcting codes, it becomes increasingly improbable that the superficial similarities between language and information structure arise by chance. Instead, we must acknowledge a deep causal link between language and information, reflecting the relationship Plotkin and Nowak claim to subsist between their fitness functions. Linguistic structure is thus a vestige of the evolutionary importance of preserving information.

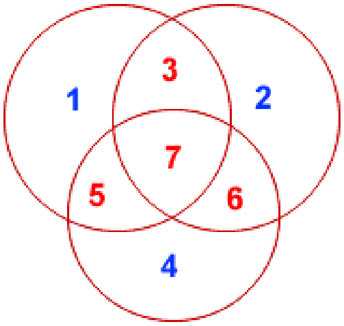
Chapter 2

Coding Strategies Compared

Before plunging head-first into linguistic analogs to codes, it is useful to undertake a rudimentary analysis of two error-correcting codes. Both codes we investigate in this chapter, the repetition code and the Hamming code, rely on redundancy, but to differing degrees. Where the repetition code relies on redundancy alone, the Hamming code structures and balances redundancy throughout the message. In our discussion we represent messages as binary strings, but the notions of redundancy and internal structure are equally pertinent to language; in some sense, then, the performance of each code points to the evolutionary benefit of converging on a comparable linguistic system.

In a repetition code, a single bit of the message is encoded as an n-tuple of that bit; for example, with a repetition block of length 3, the message [0 0 1 1 0] is written as [000 000 111 111 000]. The effectiveness of this method is dependent on the number of repetitions employed, correlating highly accurate transmission with a drastic increase in message length. Hamming codes, by contrast, rely on the insertion of a small number of 'parity bits,' which form blocks summing to 0 modulo 2 when the bits of the message are arranged in the following Venn diagram:

Figure 2.1: Configuration of Parity-Setting Bits in a Hamming Code



.

Bits 1, 2, and 4 in figure 2.1 are the parity bits, each of which is defined in relation to the three bits of the original message that fill the large circle in which it falls (Tervo 2005). The importance of parity in the sub-sections of a Hamming-encoded message highlights relationships among sub-units, a strategy perhaps analogous to the agreement between parts of speech in many modern languages. The necessity of attaining a sum of 0 modulo 2 within a given configuration also restricts the number of possible combinations, since only the combinations in which all bits are 0, all bits are 1, or half the bits are 0 and half are 1 will attain the desired total. Specifically, due to the bit overlap in the parity-setting configuration, this combination must be arranged correctly in the corresponding Venn diagram to ensure that all blocks also contain similarly acceptable combinations. Emphasizing both limitations on possible sequences of sub-units as well as their arrangement within a structured environment suggests a parallel to restricted phonetic or lexical inventories as well as rules governing linguistic structure.

2.1 Redundancy vs. Structure

We establish a baseline noise probability – the probability that any given bit will switch, for example, from 0 to 1 – of 0.1 and a standard message length of 3 bits. To compose a random message, we generate a random decimal between 0 and 1 for each bit of the message; if the decimal is greater than 0.5, we add a 1 to the message, but if the decimal is less than or equal to 0.5, we add a 0. We then encode the message using a repetition code by creating a new empty message; for each bit in the original message, we repeat that bit a specified number of times in the encoded message. We expose such a message to noise by creating a new empty message of the same length as the encoded message. For each bit in the encoded message, we again generate a random decimal between 0 and 1. If the random decimal exceeds the noise probability, we copy the encoded bit into the new message. If the noise probability exceeds the random decimal, we add the opposite bit to the new message. The decoding procedure then divides the received message into chunks of the same length as the encoding block. Whichever bit appears most often in the received chunk is selected as the most likely original bit and is added to the decoded message. If the chunk is evenly split between the two possible bits, we choose the original bit at random. The encoding and decoding functions for the Hamming code are more complex, but the message generation and noise procedures are the same. In both cases, the newly decoded message is re-encoded, re-exposed to noise, and re-decoded to become the input message for another iteration of encoding, noise exposure, and decoding. After each decoding we determine whether the original and received messages are the same in order to tabulate the code's cumulative success rate. We repeat the process 100 times in each transmission chain before assessing the performance of the code.

Throughout the simulations to be described, we measure a coding algorithm's success based on its transmission accuracy, or the percentage of transmitted messages that are correctly decoded after being exposed to noise. We do so primarily because transmission accuracy is the standard of proof in information theory, which is concerned with the identification and correction of errors rather than the communication of concepts, and also because of the success this measure has achieved in the iterated learning model (ILM). The premise of the ILM, as discussed, for example, in Kirby (2001), is that linguistic structure can arise through repeated transmission of linguistic data from adults to learners even in the absence of explicit selective pressure. This paradigm is particularly fitting for the current study because it links the central precepts of information theory to the emergence of linguistic structure.

We begin with repetition codes, since the accuracy of a repetition code depends on two factors: the length of the original message and the size of the encoding block, or the number of repetitions per bit. We encode messages of 3 and 8 bits using repetition blocks of 3 and 5 bits. The results are summarized in figure 2.2:

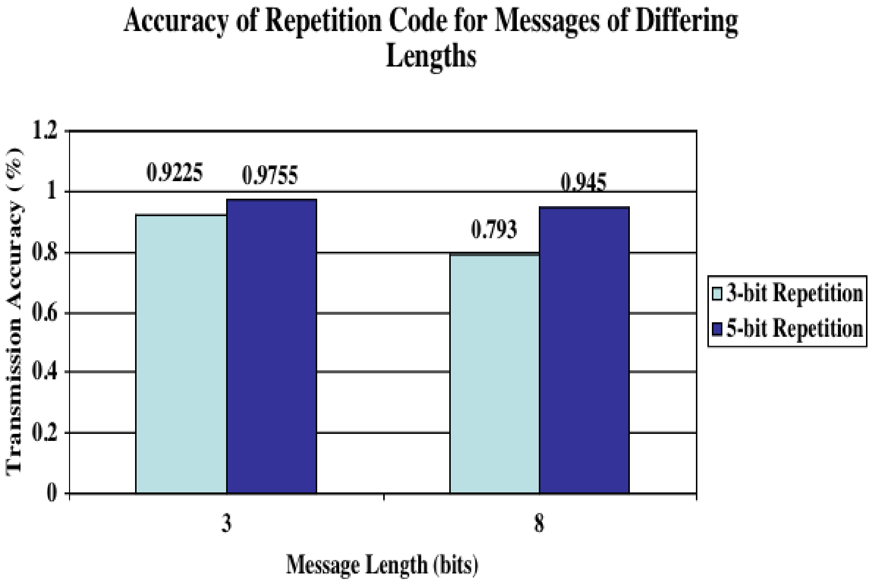


Figure 2.2: Factors Affecting the Performance of Repetition Codes

Although transmission accuracy decreases with increasing message length for both lengths of repetition block, the decrease is less dramatic with a repetition block of length 5, suggesting that longer repetition blocks can in part mitigate the decline. We perform a one-way ANOVA to determine the significance of the effect of block length on transmission accuracy. Both the main effects – the effect of message length and of block length – are significant, $F(3) = 52.47$, $p \leq 0.001$ for message length and $F(1) = 388.812$, $p \leq 0.001$ for block length, as is the interaction between block length and message length, $F(3) = 22.104$, $p \leq 0.001$. Our intuition that increasing the number of repetitions enhances the accuracy of transmission is therefore validated. For repetition codes, long repetition blocks slow the decrease in accuracy across long messages.

Unlike repetition codes, Hamming codes have only one parameter: message length. Because of the specific configuration required to set a Hamming code's parity bits, Hamming codes have two possible lengths: four bits, encoded with seven bits, or eleven bits, encoded with fifteen bits. Since no such constraints operate on repetition codes, we perform 20 runs of 100 transmissions of messages of length 4 and 11 for both types of code.

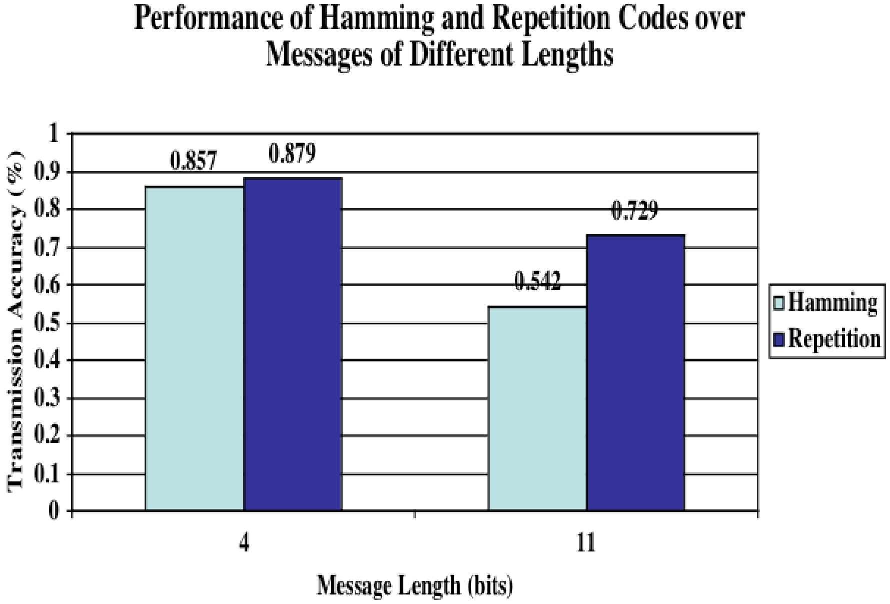


Figure 2.3: Compared Accuracy of Coding Strategies: Hamming vs. Repetition Codes

Although repetition codes appear to be more accurate overall, an independent-samples t-test reveals no significant difference between Hamming and repetition codes' accuracy for four-bit messages. For longer messages, however, the difference becomes significant, with repetition codes producing higher accuracies than Hamming codes as confirmed by a one-way ANOVA, $F(1) = 417.05$, $p \leq 0.001$.

Judged on transmission accuracy alone, repetition codes are more effective than Hamming codes as message length increases. For short messages the two are equally accurate, but the possibility of increasing the ratio of encoded to original bits – linguistically, increasing redundancy – allows the repetition code to take the lead.

2.2 More Complex Noise

From a linguistic perspective, the superiority of the repetition code seems counter-intuitive; internal structure is a more satisfying analogy to syntax, and therefore ought to be more accurate, if we consider syntax to be the result of an evolutionary process favoring transmission accuracy. Repeating every segment many times, while accurate, forces a trade-off between transmission accuracy and message length and may therefore be maladaptive with regard to cognitive constraints.

To replicate the effect of limited working memory or cognitive efficiency, we re-assess the two conventional error correcting codes under noise functions constructed to penalize increased message length. We create five new noise functions: a step-wise function, a linear function, a quadratic function, a chunking function, and a conditional or 'neighbor' function. In general, we increase the probability of flipping a given bit based on that bit's position in the message, making bits later in the message more likely to be flipped. In the step-wise function, for example, we record the position of each bit in the message, that is, the first bit is bit 0, the second is bit 1, etc, and increase the noise probability by a given increment, in this case 0.1, whenever the bit's position exceeds a multiple of three. In the linear function the noise probability is represented by the equation

$$\text{noise} = 0.05x + 0.1,$$

where x represents the position of the bit in the message. The quadratic noise probability is also a continuous function of the bit's position in the message according to the equation

$$\text{noise} = 0.004\,x^2.$$

The coefficients in both the linear and quadratic equations are arbitrary; they were chosen to lower transmission accuracy without eliminating the possibility of correctly decoding a message.

The chunking and conditional noise functions replicate conditions of the communicative environment. In conversation, a disturbance interrupting the flow of auditory data will not confine itself to the boundaries of a given message unit, but will instead affect several consecutive units. We therefore divide the message into chunks of a specified size and set a baseline noise probability; if the noise probability exceeds a randomly-generated decimal, all bits in the segment are flipped. The 'neighbor' noise function springs from a similar concept, but does not require all units in the vicinity of an interrupted unit to undergo the same amount of disruption. Instead, when exposing a message to noise we record the origin of each bit in the received message, that is, whether the 'new' bit has been flipped or is the same as the original bit. If the previous bit was flipped, the current bit's probability of flipping increases from 0.1 to 0.3. In this way we compare the effect of flipping large units simultaneously, as in the chunking function, to a probabilistic domino effect making future errors more likely when an error has already occurred.

We assess both Hamming and repetition codes under all five functions. We run two trials of the repetition code, first with an encoding block of length three and then with a block of length five. Since we aim to penalize excessively long messages, we expect transmission accuracy to decrease with longer repetition blocks. We also expect the Hamming code, which incorporates fewer redundant bits into its encoded messages, to out-perform the repetition code. We present the data below:

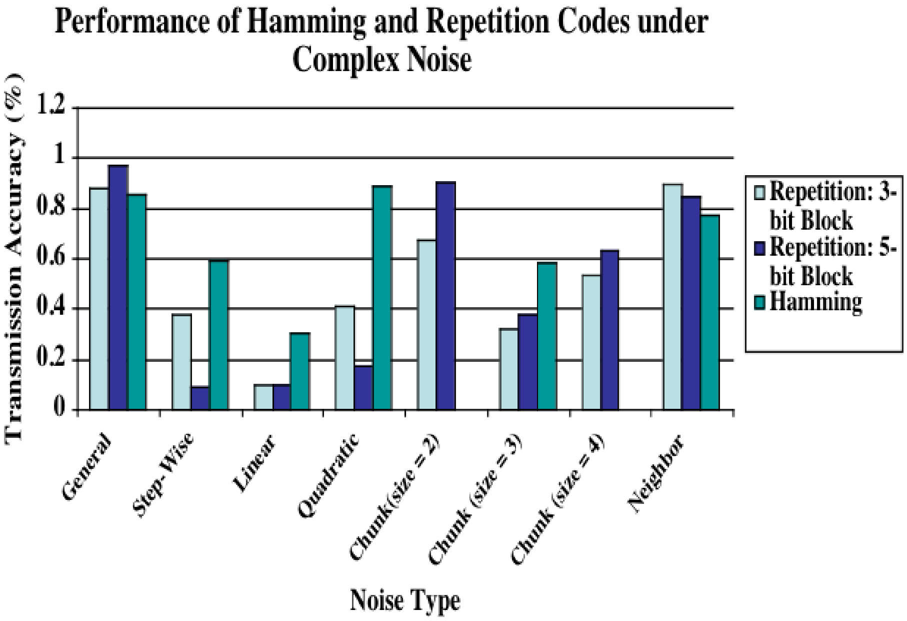


Figure 2.4: Relative Success of Hamming Codes under Complex Noise

In general, our predictions are correct; not only do the new noise functions reduce the accuracy of repetition codes substantially, transmission accuracy decreases as the length of repetition blocks increases. In fact, the linear and quadratic noise functions reduce transmission accuracy so dramatically that it is necessary to alter the function coefficients to 0.02 and 0.002, respectively, to avoid obtaining transmission accuracies of zero.

The chunk noise method, however, exhibits the reverse of the overall negative correlation between accuracy and repeated block length. This contradictory trend is less surprising than it may at first appear, since the chunk function does not penalize message length – it broadens the effect of the noise, but does not map noise probability directly to increased message length. If the number of repetitions employed by the encoding function is significantly larger than the chunk size, the encoding function can remain impervious to the effect of the chunk noise function. The effect of larger encoding block is reduced by employing a larger chunk size. Similarly, repetition codes out-perform Hamming codes under the neighbor noise function, which, like the chunk function, does not penalize increasing message length.

Under the noise functions designed to penalize message length, however, Hamming codes are less susceptible to increased error. In order to compare Hamming codes to repetition codes under the chunk noise function, we are restricted to a chunk size of three and a message length of eleven, because in other conditions the length of the Hamming-encoded message does not divide into even chunks and therefore cannot be analyzed appropriately. In comparison to a comparable run using a repetition code, however, the Hamming code is more accurate.

When noise is represented as an invariant probability, repetition trumps internal structure, but that advantage disappears when messages are subjected to more sophisticated noise functions. Although simple and structured redundancy are equally advantageous for short messages, structure becomes crucial for long messages, especially when communication is evaluated under varying noise types. The viability of internal structure as a coding strategy lends credence to the theory that linguistic structure may have a similar information-preserving function.

Chapter 3

Language-Like Codes

3.1 A Linguistic Message-Construction Algorithm

Internal structure, as demonstrated by the success of Hamming codes under complex noise functions, facilitates the identification and correction of transmission errors. Although the Hamming code's internal structure is roughly comparable to syntactic structure, the organization employed in the Hamming code has no specific linguistic correlate. In order to buttress the theory that linguistic structure preserves information under noise, it is crucial to demonstrate that algorithms designed to replicate linguistic features can perform as well as, if not better than, error-correcting codes.

We begin with a message construction algorithm encapsulating a rudimentary lexical or phonemic system. As previously discussed, there is substantial evidence in the literature that maximal distinctiveness among signal units is fundamental to linguistic communication. Although these simulations examine communicative accuracy, or the ability of two communicative agents to converge on a mutually comprehensible system, and we investigate transmission accuracy, or conservation of the form of a signal, it is reasonable to imagine that mutually distinct segments would also be beneficial when maximizing transmission accuracy. We therefore seek to determine the benefit of restricting input to a set of fixed combinatorial segments; in keeping with the results of Liljencrants and Lindlbom (1972), de Boer (2000), Zuidema and Westermann (2003), and Zuidema and de Boer (2009), we develop a set of maximally distinct binary segments. We measure distinctiveness as distance in number of digits that must flip in order to transform one segment into another, a rudimentary re-working of the informational measure Hamming distance; each of the three segments is at least two, if not three, flips distant from the other two. The segments are: [1 0 0 0], [0 1 0 1] [0 0 1 1].

We employ the same transmission chain methodology as in the previous simulations, starting with a randomly-generated message comprised of the candidate subsegments which is then exposed to noise as before and decoded to produce a new message. When the message is exposed to noise the segment boundaries are removed, so the noise function operates on every bit rather than every segment. The received message is decoded as follows: the series of bits is divided into chunks of the same length as the original segments and the number of positions in which the received segment differs from each pre-specified segment is calculated. The pre-specified segment that differs in the fewest positions from the new segment is selected as the most likely original segment and is added to the new message. The new message is then subjected to noise, decoded, and used as the new message in the next iteration. All tests consist of 20 runs of 100 iterations per run. We assess the transmission accuracy of different message lengths under different noise functions. First we present the baseline results, observed with a standard noise probability of 0.1.

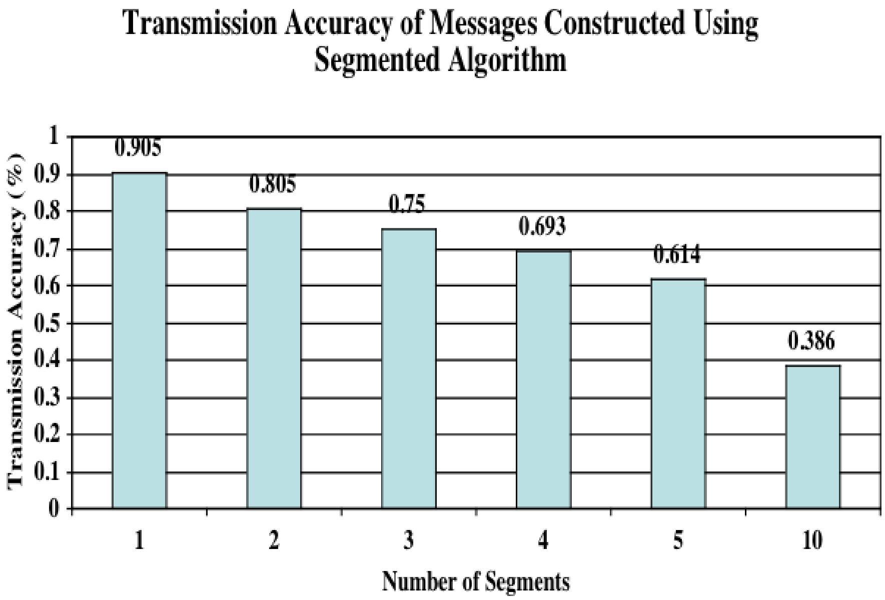


Figure 3.1: Effect of Long Messages on Accuracy of Segmented Algorithm

As expected, cumulative transmission accuracy decreases with increasing message length. Compared to Hamming and repetition codes, however, the segmented construction method remains relatively impervious to noise for much longer messages. The repetition code is assessed using an encoding block of three bits. Because the segments are a pre-defined length, it is difficult to compare transmission accuracies directly among the three types of code investigated; it is possible, however, to extrapolate from the relevant figures. We present the comparative data below, employing message bits as the unit of analysis rather than segments to facilitate comparison among the three coding methods:

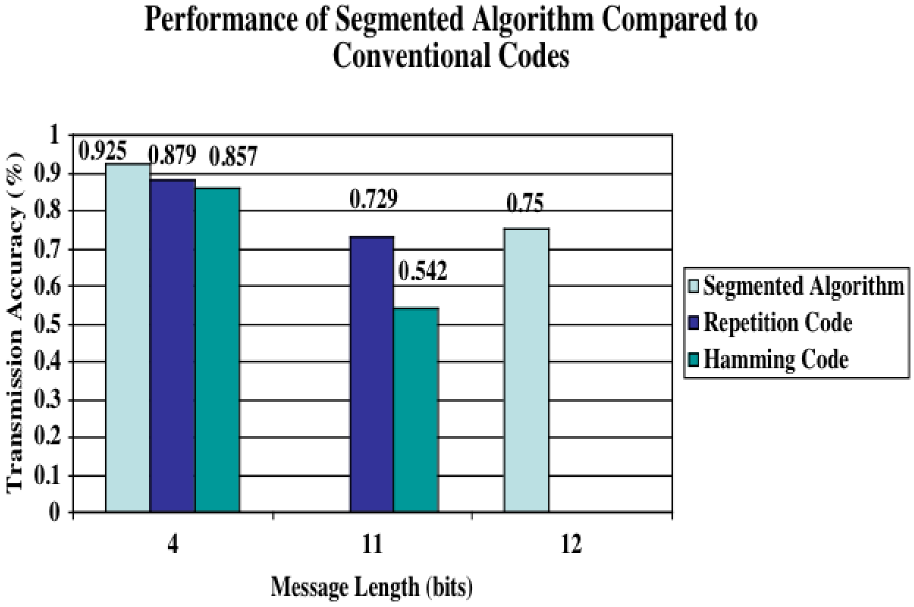


Figure 3.2: Improvement in Transmission Accuracy with Segmented Algorithm Compared to Conventional Codes

The segmented construction method appears to perform better at all message lengths than either the Hamming or repetition codes under standard noise conditions. It is difficult to verify the difference statistically, given the constraints on message length for both the segmented algorithm and the Hamming code, but the data strongly suggest that the segmented algorithm is at least as accurate as the error-correcting codes under general noise. Given that general noise, while an important basis for comparison, does not reflect varieties of noise occurring in the communicative environment, we subject the segmented algorithm to the five complex noise functions developed in the previous chapter, with the following results:

Table 3.1: Transmission Accuracy of Segmented Algorithm under Complex Noise Functions

Noise Type	Number of Segments	Accuracy	Standard Deviation
Step-Wise	2	0.562	0.048
	3	0.244	0.039
Linear	2	0.335	0.048
	3	0.055	0.027
Quadratic	2	0.793	0.055
	3	0.336	0.062
Chunk (size = 3)	2	0.98	0
	3	0.679	0.041
Chunk (size = 4)	2	0.81	0.045
	3	0.738	0.043
Neighbor	2	0.769	0.038
	3	0.672	0.043

The high performance of the segmented construction method remains undiminished under more complex noise functions, most of which are designed to penalize increasing message length. Given that the segmented algorithm necessarily involves the production of longer messages – a message with three segments contains twelve digits, which is longer than the longest Hamming code investigated – its relatively high performance indicates that forms of internal structure may compensate for the higher probability of noise concurrent with longer messages. We compare accuracy figures to equivalent runs of repetition and Hamming codes in the table below; again, due to the constraints on message length imposed by the segmented message construction, we must extrapolate from the given data.

Table 3.2: Transmission Accuracy of Segmented Algorithm Compared to Conventional Error-Correcting Codes

		Transmission Accuracy		
Noise Type	Number of Bits	Segmented	Repetition	Hamming
Step-Wise	4		0.376	0.594
	8	0.562		
Linear	4		0.0975	0.308
	8	0.335		
Quadratic	4		0.4135	0.891
	8	0.793		
Chunk (size=3)	4		0.319	0.58
	8	0.98		
Neighbor	11		0.736	0.475
	12	0.672		

Comparison of the relevant figures indicates that the segmented construction method performs better than either error-correcting code under each type of complex noise, underlining the importance of fixed symbol sets in ‘noisy’ communication.

3.2 A Statistical Message-Construction Algorithm

In human communication, the items that comprise a set of discrete linguistic elements are rarely used with uniform frequency. Some linguistic combinations occur more frequently than others, an idea Harris highlighted in his emphasis on deviation from equiprobability. Recent evidence augments the importance of this perceptual strategy, as discussed earlier in relation to both first language acquisition and mathematical approximation of natural language. Awareness of the statistical relationships that subsist among lexical or phonetic items should thus enhance our segmented message construction algorithm.

In the next simulation, we implement a transition matrix, that is, a matrix specifying the probability of transition from the current segment to each of the possible segments in the segment inventory, in the message-generation and decoding processes. The set of permissible segments remains unchanged. When creating a message, the first segment is chosen at random, but the subsequent segments are selected stochastically using the transition matrix. As before, the messages are exposed to noise and the resulting message is decoded. The statistical decoding function counts the number of bits the segment in question has in common with each of the predetermined segments, then multiplies the number of common bits by the probability that that segment would follow the previous segment to be decoded, essentially scaling the similarity between the received segment and each permissible segment by the permissible segment's probability of appearing in that sequence. The segment with the highest scaled score is selected as the most likely original segment. We develop a quasi-random matrix designed to highlight differences in conditional probability distributions among segments. In this matrix, no segment has a uniform distribution, so there are most and least probable transitions for each segment, and no segment is most likely to transition to itself, underlining what we imagine to be a potentially useful strategy for discriminating segment recurrence patterns. We present the matrix in figure 3.3.

Figure 3.3: Transition Matrix Employed in Statistical Message Construction Algorithm

	0	1	2
0	0.25	0.25	0.5
1	0.7	0.2	0.1
2	0.3	0.6	0.1

We tabulate the algorithm's accuracy over 75 runs of 1000 transmissions under each of our six noise functions. The increased structure present in the message appears to correlate to a relatively high level of transmission accuracy, as indicated in table 3.3, where we compare accuracies attained using the segmented and statistical construction algorithms. We also include the results of independent-samples t-tests performed between each pair of corresponding conditions.[1]

[1] A multi-way ANOVA would be a more appropriate test, but compatibility with statistical software restricted us to the analytical techniques available in the Python statistical package. This is the case for all analyses presented in later chapters.

Table 3.3: Performance of Statistical vs. Segmented Algorithms under Complex Noise

		Construction Method		t-test Output	
Noise Type	Number of Segments	Segmented	Statistical	t	p
General	2	0.805	0.871	40.54	4.97E^{-82}
	3	0.749	0.843	45.55	5.68E^{-89}
Step-Wise	2	0.562	0.74	30.66	5.38E^{-66}
	3	0.244	0.552	63.64	2.09E^{-109}
Linear	2	0.335	0.545	36.19	2.14E^{-75}
	3	0.055	0.269	57.338	5.76E^{-103}
Quadratic	2	0.793	0.947	24.21	2.54E^{-53}
	3	0.336	0.7016	48.78	4.33E^{-93}
Chunk (size=3)	2	0.98	0.998	-103.95	3.09E^{-140}
	3	0.679	0.769	17.63	3.46E^{-38}
Chunk (size = 4)	2	0.809	0.804	-1.213	0.227, n.s.
	3	0.738	0.7155	-4.38	2.25E^{-5}
Neighbor	2	0.769	0.813	9.554	3.82E^{-17}
	3	0.672	0.7593	16.61	1.27E^{-35}

In most cases, as is evident from table 3.3, the statistical method edges out the segmented method. In fact, only one of the independent-samples t-tests, that comparing the statistical to segmented method under chunk noise with a chunk size of four and a message length of two segments, is not significant. It is also in the chunk noise condition that we observe the reverse of the positive trend evident in all other comparisons; with a chunk size of four, the segmented method out-performs the statistical method, which is superior in all other comparisons. A chunk of length four is the same length as the possible segments and would thus switch every bit in the segment, making it difficult to identify the correct original segment; since the viability of the statistical method depends in part on correctly identifying the original segment, the addition of probabilistic information is not useful in this condition. We also note that in general the difference between transmission accuracies is more significant when comparing performance under messages of length three, suggesting that the statistical method also mitigates the nugatory effect of increased message length.

It is unsurprising that the statistical algorithm improves on the segmented algorithm; the statistical algorithm reaps all the benefits of a predetermined set of acceptable combinations while incorporating an additional source of information. Crucially, out-performing the segmented method means the statistical method also out-performs both Hamming and repetition codes, which are less accurate across the board than the segmented algorithm. The fundamental role of probabilistic analysis in language perception and approximation, coupled with its success as a coding mechanism, points to a strong link between language and error-correcting codes. As we have seen, construction algorithms designed to replicate basic linguistic features are more accurate than abstract codes. The potential for a positive correlation between an algorithm's fit as a linguistic approximation and its efficacy as an error-correcting code strongly supports the hypothesis that linguistic structure reflects not only the form but also the function of information theory.

Chapter 4

Evolving Language-Like Codes

While the segmented and statistical construction methods indicate that linguistic structure can approximate the effects of error-correcting codes, it is unfair to compare them directly to linguistic features. These algorithms, although designed to imitate naturally-occurring linguistic structures, are explicitly constructed to maximize communicative accuracy. In the segmented construction method, for example, we actively selected segments based on their mutual distinctiveness, a pressure recognized as crucial. It is thus possible to attribute their success to design properties, like maximal distinctiveness, rather than to the strict adaptiveness of the methods themselves as coding strategies.

In an evolutionary framework, it is not sufficient merely to demonstrate that a method of message construction using a specific dictionary of permissible segments successfully preserves information under varying degrees of noise. It is necessary to show that natural evolutionary processes can produce the same high levels of communicative accuracy. We therefore adopt the same basic framework – a method of message construction relying on a dictionary of predetermined segments – but develop simulations to evolve optimal segment dictionaries and optimal transition matrices through random variation.

The new simulations reflect a broad trend in computational research in language evolution, epitomized in Zuidema and Westermann (2003). As previously discussed, Zuidema and Westermann formulate an expression for the communicative success of a system of form-meaning mappings in a population of simulated agents order to determine which conditions produce optimal results (2003: 390). Each agent in the simulation begins with two random stochastic matrices, one representing signal production, the other representing signal reception. To this population Zuidema and Westermann apply a hill-climbing algorithm: assess the communicative fitness of the original population; make a random alteration to a random agent's production or reception matrix; re-assess communicative fitness; if the change improves communicative fitness, set the new matrix as the baseline and perform further random refinements; if not, return to the previous state, and repeat the process until communicative fitness no longer improves (ibid). Based on this algorithm, Zuidema and Westermann investigate the optimal systems emerging under diverse pressures, from noisy signaling to semantic similarity, without imposing assumptions about optimal structures.

4.1 Evolving Optimal Dictionaries

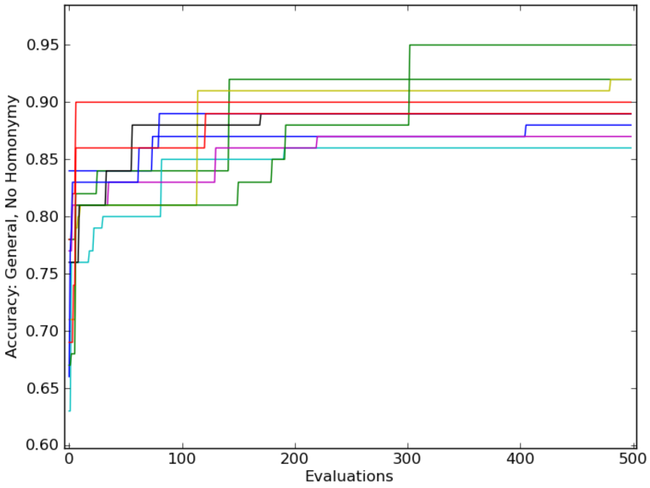
We apply a similar hill-climbing method to evolve segment dictionaries that optimize transmission accuracy. We begin with a randomly-generated set of four-bit segments, which we then expose to noise and decode iteratively using the noise and decoding functions described in earlier chapters, recording accuracies. A random segment from the dictionary is selected to be replaced by another randomly-generated segment and the new dictionary's accuracy is assessed. If the new dictionary's accuracy is greater than the previous dictionary's, we repeat the process with the new dictionary; if not, we return to the previous dictionary and make another random modification. The number of evaluations required for convergence varies by noise condition – in general, noise conditions that constrain transmission accuracy more strictly require more evaluations.

In our first trial, all of the systems investigated under all six noise functions attain perfect accuracy. Upon examining the resultant dictionaries for these highly successful optimization procedures, however, we immediately observe a striking and ubiquitous effect of homonymy; most dictionaries contain repeated segments, and some contain only one segment repeated three times. From an error-correcting code perspective, highly repetitive or homonymous dictionaries are indeed optimal, since reducing the number of possible segments increases the probability of selecting the correct segment at random. Experimental evidence from the iterated learning model also supports the primacy of homonymy as a linguistic predisposition (see Kirby, Cornish, and Smith 2008). Although a linguistic system with a single segment in its phonetic or lexical inventory is guaranteed to be transmitted accurately, such a system can convey very few messages. Since homonymy is maladaptive as a basis for communication, we discard these results.

4.1.1 Results: Non-Homonymous Dictionaries

We introduce a homonymy filter into the dictionary generation algorithm by designing a function that compares each segment of the dictionary to each other segment; if two segments are the same, one is removed and a new randomly-generated segment is added, and the segment-by-segment comparison continues until no identical segments are identified. Every random modification of the original dictionary is filtered for homonymy to ensure that repetitions do not arise through random variation. We perform 75 runs of each noise condition, with the number of evaluations varying by condition as before. When every segment in a successful dictionary must be unique, the success rate attained even in an evolutionary setting decreases substantially. Nonetheless, evolved segment dictionaries demonstrate an improvement on their random counterparts, as is evident from the postive slope of the representative evolutionary trajectory we present in figure 4.1.

Figure 4.1: Increase in Transmission Accuracy throughout Evolution of Segment Dictionaries under General Noise



Ideally, the optimized dictionaries would improve not only on the performance of random initial dictionaries but also on the accuracies obtained using a dictionary designed to maximize inter-segment distinctiveness. We compare final accuracies from the designed and evolved dictionary simulations in figure 4.2.

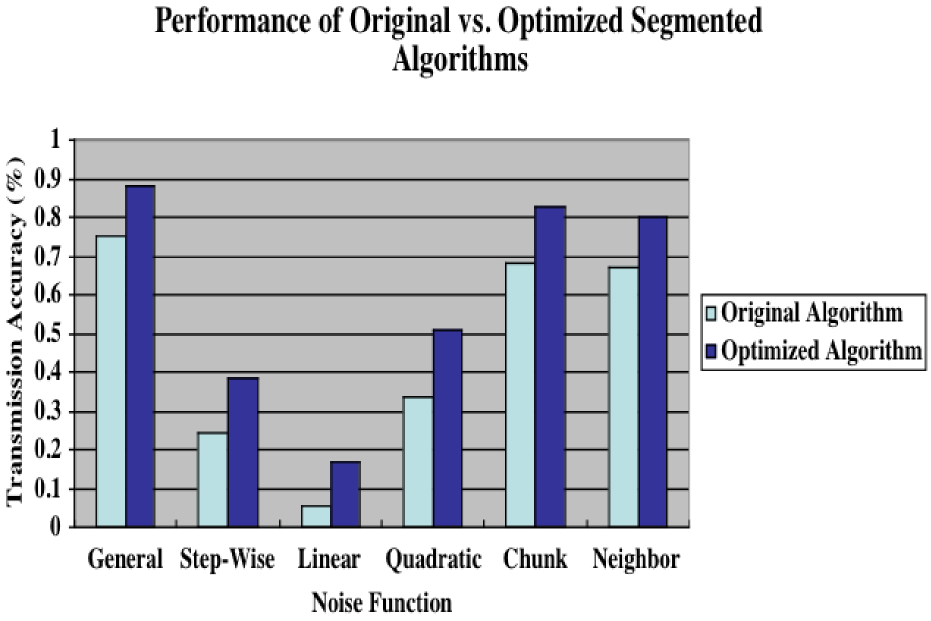


Figure 4.2: Improvement in Transmission Accuracy with Evolved Segment Dictionaries

It is clear, even from a casual inspection of the accuracy values, that the optimized segment dictionaries are more effective. We perform an independent-samples t-test comparing the across-the-board accuracy values, grouping all noise conditions together, to determine whether the apparent difference between accuracy values is spurious. Confirming our intuition, the difference between sample means is indeed significant, $t(118) = -4.029$, $p = 7.53\text{E}^{-5}$. The evolved segment dictionaries mark a substantive improvement on the original dictionary.

4.1.2 Assessing Segment Distinctiveness

The superiority of the randomly-evolved dictionaries calls the importance of maximal distinctiveness into question, given the primacy of maximal distinctiveness in the construction of the original segment dictionary. We therefore attempt to quantify the structure of the optimized dictionaries through several parameters: the maximum distance, in bit-flips, between segments in a given dictionary, the minimum inter-segment distance, whether all segments in the dictionary have the same number of 1s, and whether all segments in the dictionary have different numbers of 1s. These metrics highlight the extent to which segments in a dictionary are mutually distinct; the maximum and minimum distances, for example, specify the range of bits any two segments in the dictionary have in common, so if the maximum and minimum inter-segment distances are both high, there is a relatively high mutual inter-segment distance obtaining for that dictionary. The composition metrics approach mutual distinctiveness from a different perspective. It is possible that a dictionary in which each segment has the same number of 1s (and thus the same number of 0s) would be easier to decode accurately, since any segment with a different number of 1s could not belong to the dictionary. Similarly, a dictionary in which each segment has a different number of 1s could reduce the possibility of mistaking one permissible segment for another. We investigate the extent to which these distance metrics are correlated to transmission accuracy. In keeping with Liljencrants and Lindblom (1972) and Zuidema and Westermann (2003), we predict that accuracy and maximum inter-segment distance, minimum inter-segment distance, and whether all segments have different numbers of 1s, our indicators of distinctiveness, should be positively correlated.

Contrary to expectation, however, correlational statistics reveal no systematic relationships between structural features and transmission accuracy across noise conditions. In fact, of the 24 correlations performed, only one is significant at the $\alpha = 0.05$ level. The relative ambivalence manifested in the overwhelming number of non-significant correlations strongly indicates the absence of a consistent relationship between dictionary structure and transmission accuracy.

4.1.3 Larger Dictionaries, More Structure?

Unlike the systems investigated by Liljencrants and Lindblom and Zuidema and Westermann, our segmented message construction algorithm does not evince a preference for maximally distinct segments. It is possible, however, that the dissociation of accuracy and structure reflects the wide range of combinations available for the construction of a three-segment dictionary; perhaps, then, increasing the number of segments in the dictionary will increase the pressure for distinctiveness, highlighting the relevance of structural factors. To test this hypothesis, we repeat the hill-climbing optimization procedure using randomly-generated four- and six-segment dictionaries. As before, we perform 75 runs of each noise condition for each dictionary size condition, with representative trajectories below.

Figure 4.3: Evolution of Transmission Accuracy when Optimizing Four-Segment Dictionaries under General Noise

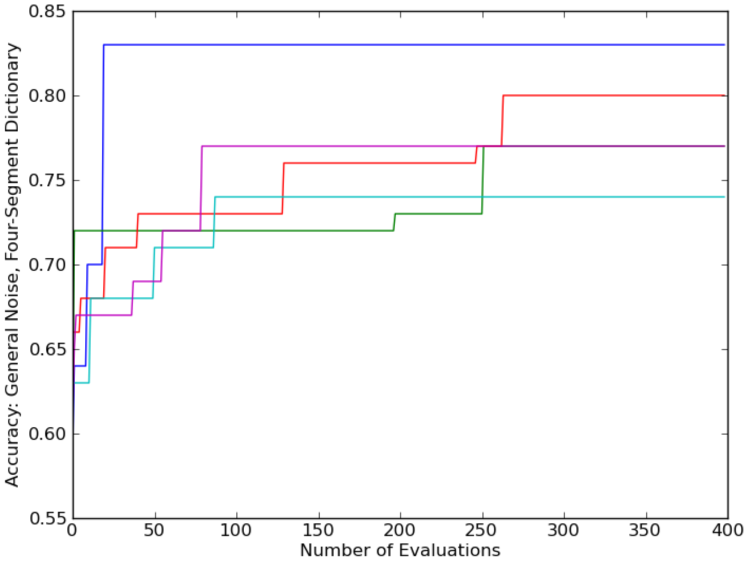
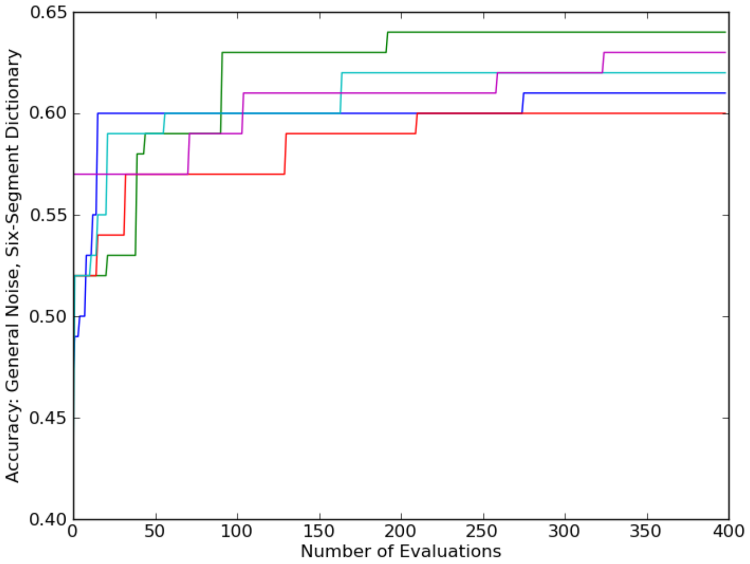


Figure 4.4: Evolution of Transmission Accuracy when Optimizing Six-Segment Dictionaries under General Noise



Transmission accuracy appears to decrease with increasing dictionary size. The trend becomes more concrete on direct inspection of the average accuracy attained in each condition under each noise function:

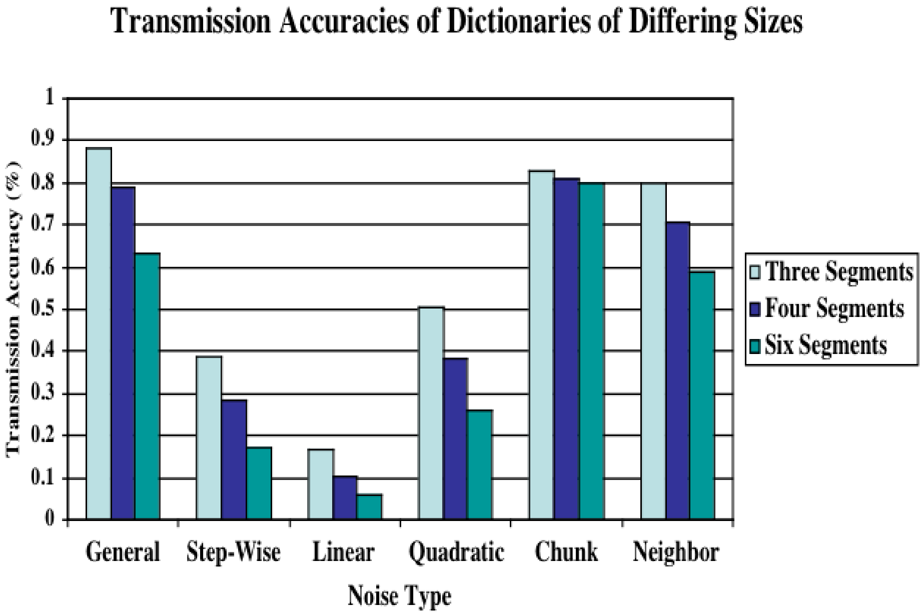


Figure 4.5: Decrease in Transmission Accuracy with Increasing Dictionary Size

We assess the apparent decrease in accuracy with increasing dictionary size using six one-way ANOVAs in which we hold noise type constant and compare accuracies by dictionary size. The results strongly support a significant difference in transmission accuracy between length conditions of the same noise type. We summarize the results in table 4.1.

Table 4.1: Parameters of One-Way ANOVAs Comparing Transmission Accuracies of Dictionaries of Different Sizes

Noise Type	F	p
General	1965.08	6.56E^{-142}
Step-Wise	2365.29	2.08E^{-150}
Linear	752.67	1.25E^{-99}
Quadratic	1821.8	1.83E^{-138}
Chunk	45.23	3.34E^{-17}
Neighbor	1674.31	1.23E^{-134}

We also perform post-hoc independent-samples t-tests comparing accuracies of consecutive dictionary sizes in each noise condition to determine whether these results reflect a consistent trend; all post-hoc comparisons are also significant, $p < 0.015$. The ANOVAs and independent-samples t-tests confirm the hypothesis that transmission accuracy decreases with increasing dictionary size.

The most parsimonious explanation for this striking decline is that increasing the number of segments in the dictionary reduces the probability of selecting the correct segment by chance. Once again, the system emphasizes its preference for homonymy: the fewer possible segments, the more accurate the transmission.

Although homonymy appears to be the primary factor affecting transmission accuracy in these communicative systems, it would be premature to attribute the accuracy decline to this factor alone. We therefore apply the four metrics discussed above – maximum and minimum inter-segment distance as well as whether all segments have the same or different numbers of 1s – to the dictionaries produced in the four- and six-segment dictionary conditions. As before, we perform Pearson correlations between each of these metrics and transmission accuracy for each condition.

As in the three-segment dictionary condition, in the four-segment dictionary condition there are few significant correlations between transmission accuracy and the four structural parameters. Again, there is no evidence of systematic relationships between structural features and communicative success. The complete absence of significant correlations in the six-segment condition confirms the dissociation of structure from transmission accuracy.

Although none of the structural attributes is a significant predictor of transmission accuracy, further statistical analyses reveal strong relationships among these attributes and the conditions under which they emerge. We seek to determine whether, holding noise type constant, the structural metrics vary depending on the size of the dictionary. For each structural metric we perform a one-way ANOVA comparing values observed in dictionaries of different lengths produced under the same noise condition. Both minimum and maximum inter-segment distance, for example, vary significantly by dictionary size when noise type is held constant. All of these comparisons are highly significant, $p << 0.0001$, and suggest a natural consequence of the homonymy bias: as more segments are added to the dictionary, not only does the probability of selecting the correct segment at random decrease, but the probability that the segments will be structurally distinct also decreases.

We observe a similar effect for the two composition metrics, whether all segments in a given dictionary have the same or different numbers of 1s. When comparing the number of dictionaries in which all segments have the same number of 1s, there is no significant difference by dictionary size when noise type is held constant, but the reverse is true for dictionaries in which all segments have different numbers of 1s. Not only are all comparisons by dictionary size significant, $p << 0.001$ in all cases, but in most cases the post-hoc tests reveal significant differences between the composition percentages for each dictionary length. There are only two post-hoc comparisons for which sample means are indistinguishable: under both quadratic and neighbor noise, three-segment and four-segment dictionaries have roughly similar compositions. Both of these comparisons are significant at the $\alpha = 0.05$ level, but applying the Bonferroni correction for post-hoc tests requires us to reduce significance to $\alpha = 0.017$; the p-values in question remain marginally significant even at this level, with $p = 0.03$ and $p = 0.024$, respectively.

Of course, as the number of segments increases, the number of options for non-homonymous combinations decreases, reducing opportunities to maximize inter-segment distance or rely on different ratios of 1s to 0s to distinguish one segment from another. Statistical comparison to the structural attributes of randomly-generated dictionaries reveals no significant difference between random and optimized dictionaries; this significant effect, then, springs from the combinatorial properties of segment dictionaries rather than the optimization process.

4.2 Discussion

Across dictionary lengths, when evolving segment inventories solely for transmission accuracy under noise, segment distinctiveness is irrelevant; homonymy, or near-homonymy, is the primary predictor of transmission accuracy. Our systems' near ambivalence with respect to maximal distinctiveness contrasts sharply with research on the emergence of phonemic and lexical inventories through communication games. Why, then, do our systems respond differently?

It is important to recall that the results on maximal distinctiveness previously discussed arose in the context of simulated communication games in which agents attempted to match each others' productions. This simulation is not a communication game; it is a straight-forward assessment of cumulative transmission accuracy. It is therefore conceivable that intentional coordination between agents makes maximal distinctiveness critical. When treating language as a code, however, maximal distinctiveness is not crucial.

Prioritizing transmission accuracy does not create maximally distinct systems; instead, it generates minimally distinct systems. Counter-intuitive as this result appears, it mirrors experimental findings from the iterated learning paradigm. In Kirby, Cornish, and Smith (2008), experimenters trained participants on a set of unique words referring to colored objects in motion. The participants were then tested on a set of objects, some of which they had not seen, and their responses were used as the training material for the next generation, allowing the experimenters to observe the emergence of systematicity and compositionality in the initially random, holistic language (Kirby, Cornish, and Smith 2008: 10682). The first set of transmission chains exhibited a linear decrease in transmission error across generations, parallelling the increase in transmission accuracy obtained in our dictionary optimization simulations, concomitant with a decrease in the number of distinct strings present in the lexicon – by the final generation, the number of distinct strings had dropped from 27 to between two and five (op. cit: 10683). Human participants, then, evince a strong preference for homonymous systems, and it is only when homonymy is actively removed from the language that structure emerges (op. cit: 10684). Our results, however contradictory they appear in comparison to Liljencrants and Lindblom (1972), de

Boer (2000), Zuidema and de Boer (2009) or Zuidema and Westermann (2003), reflect a naturally-occurring phenomenon.

Coupled with the crucial difference in simulation structure noted above, the unimportance of maximal distinctiveness and the concomitant dominance of homonymy delimit the influence of transmission on communication. While an environment favoring transmission accuracy – fidelity with respect to the signal alone, without considering the meaning – favors finite phonemic or lexical inventories as a way of restricting the signal space, it is ambivalent with respect to the structure of the signal space. It is only when imitation, inter-agent interaction, and signal-meaning coordination enter the picture that structural features become critical.

Chapter 5

Evolving Optimal Transition Matrices

If evolved segment dictionaries out-perform purpose-built dictionaries, evolved transition matrices should also out-perform intentionally constructed matrices. Further, based on the accuracy disparity between the segmented and statistical message construction algorithms evaluated earlier, an optimal transition matrix should improve accuracy even when employing an optimal segment dictionary.

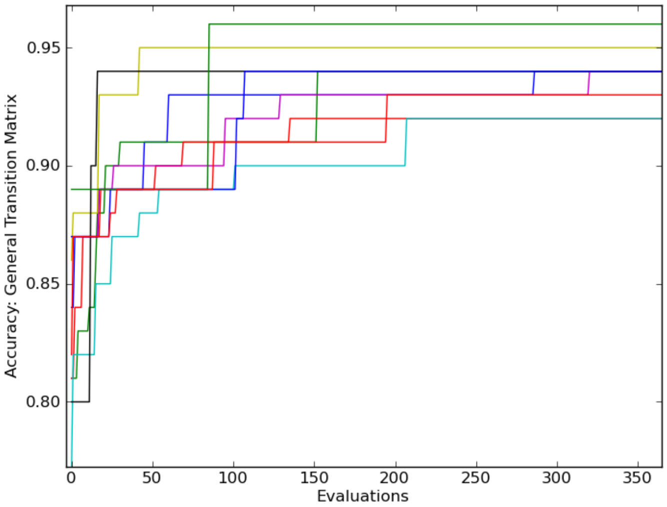
In this simulation, we again employ a hill-climbing optimization procedure to evolve a transition matrix for maximal transmission accuracy. We will discuss two main conditions, one employing a randomly-generated segment dictionary that is common across noise types, and one relying instead on successful dictionaries emerging from the previous simulation. In the optimal condition we select the segment dictionary that achieved the highest accuracy score in each noise condition and use it as the dictionary in the corresponding condition of the current simulation. We employ the same decoding function as discussed in chapter 3.

If we wish to replicate the refinement of language under natural selection according to realistic, linguistic-like cognitive processes, one could argue that it would be desirable to evolve the dictionary and transition matrix simultaneously to ensure that they are compatible. Such a simulation is beyond our technical capacity, but simultaneous optimization over both dictionaries and transition matrices also poses a logical conundrum. It is impossible to converge on a stable transition matrix without first fixing a segment inventory, since a transition matrix should reflect accumulated knowledge about co-occurrences in a fixed lexical or phonemic system. In a fluctuating system, no such accumulated knowledge is possible, so we initialize our simulation with a stable segment dictionary.

5.1 Condition I: Random Segment Dictionary

For the first set of simulations the dictionary is stable across noise conditions; the accuracy scores in these conditions should therefore reflect both the fit between the dictionary and the noise condition as well as any change in accuracy arising from the use of a transition matrix. In these runs, we employ a fixed segment dictionary of [0, 0, 0, 1], [0, 1, 0, 1], [1, 0, 1, 1]. We expect the optima obtained in this way to vary as a function of the performance of the segment dictionary under each type of noise, as demonstrated in previous simulations. We first present a representative graph of the matrices' evolutionary trajectories.

Figure 5.1: Evolution of Transmission Accuracy when Optimizing Transition Matrices Based on Random Segment Dictionaries



Again, the positive slope of the trajectories indicates improvement over the performance expected with a random segment dictionary. Having already investigated factors affecting the performance of segment dictionaries, we focus on the potential for improved accuracy when using a randomly-evolved matrix as opposed to an intentionally-constructed matrix. We compare the original and optimized transition matrix accuracies in figure 5.1.

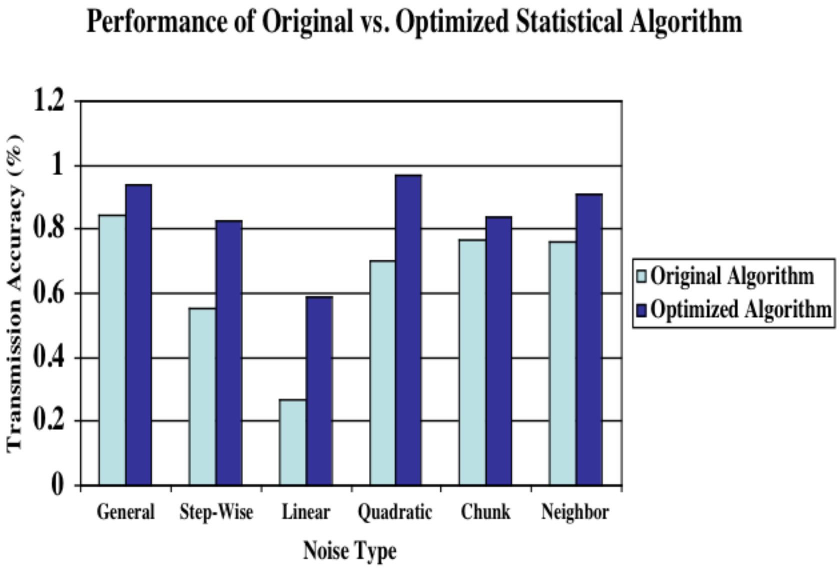


Figure 5.2: Improvement in Transmission Accuracy when Using Evolved Transition Matrices

To determine whether the apparent difference in accuracy between the two conditions is spurious, we compare accuracies attained in each noise condition and present the resulting t-test statistics below.

Table 5.1: Independent-Samples t-test Statistics Comparing Original and Optimized Statistical Algorithms

Noise Type	t	p
General	-24.535	$5.125E^{-54}$
Step-Wise	-79.45	$2.88E^{-123}$
Linear	-43.83	$1.17E^{-86}1$
Quadratic	-42.78	$3.28E^{-85}$
Chunk	-31.61	$1.06E^{-67}$
Neighbor	-55.55	$5.1E^{-101}$

As indicated in table 5.1, the evolved matrices outperform their intentionally-designed counterparts. Where the statistical method alone indicates that probabilistic analysis could improve communicative accuracy, producing higher accuracy values than conventional error-correcting codes, here we substantiate the claim that the type of matrix employed can further increase transmission accuracy even when using a sub-optimal segment dictionary.

5.1.1 Matrix Structure

It remains to investigate the structural properties of the resulting matrices. Our analysis specifically targets the amount of information encoded in the matrix through a distance analysis. We assume that the probability distributions in a useful matrix will diverge substantially from the uniform distribution, since in a uniform distribution all sequences are equally probable. We measure divergence from the uniform distribution by Euclidean distance; treating each distribution as a point in three-dimensional space, we calculate the distance between corresponding rows in two matrices, and sum the row-by-row distances to obtain a measure of the overall distance from a uniform matrix. In order to obtain a fuller picture of the factors at work in shaping a successful transition matrix, we also assess the Euclidean distance between each matrix resulting from the optimization procedure and an identity matrix of the same dimensions. This matrix represents a system that is similar to a homonymous dictionary, since it is impossible to escape from the initial state.

From a linguistic perspective, then, the identity and uniform matrices encapsulate unhelpful communicative strategies. Identity and unfiormity also represent the lower and upper bounds on Shannon entropy, defined as the amount of information or uncertainty in an information system. Shannon defines relative entropy as the ratio of a system's entropy to the entropy of a maximally uncertain system of the same dimensions (1948: 24), so we treat a matrix's distance from the uniform distribution as an approximation of its relative entropy. Neither the upper nor lower bound on Shannon entropy represents a communicatively ideal system, so we expect the resulting matrices to avoid both prototypes.

We also compare the evolved matrices to a matrix representing a potentially useful strategy: presenting the segments in a fixed order. We call this matrix the deterministic matrix and form it by shuffling the rows of the identity matrix as follows:

$$\begin{matrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{matrix}$$

Suppose we begin by selecting segment two. Segment three must always follow segment two, and segment one must always follow segment three, so the resulting pattern is 231231231.... If such a deterministic ordering is applied, it is easy to decode a message even under severe noise because correctly translating a single segment would make it possible to reconstruct the entire message. We compare the average Euclidean distances between the evolved matrices and the uniform, identity, and deterministic matrices in table 5.2.

Table 5.2: Euclidean Distance between Optimized Matrices and Three Canonical Matrices

Noise Type	Uniform Matrix	Identity Matrix	Deterministic Matrix
General	1.05	3.56	3.83
Std. Dev.	0.219	0.457	0.32
Step-Wise	1.323	3.522	3.59
Std. Dev.	0.427	0.665	0.674
Linear	1.045	3.194	3.882
Std. Dev.	0.295	0.34	0.44
Quadratic	1.25	3.922	3.158
Std. Dev.	0.223	0.325	0.603
Chunk	1.111	3.47	3.536
Std. Dev.	0.267	0.57	0.624
Neighbor	1.2	3.688	3.673
Std. Dev.	0.17	0.447	0.388

The average distances are fairly consistent across conditions. All matrices tend to prefer the uniform prototype to the identity or deterministic forms, pointing to a high level of uncertainty and therefore a relatively low level of structure. In linguistic terms, the emphasis on uncertainty over identity or determinism implies a preference for linguistic systems in which many different sequences are possible. While these values offer insight into the structure of the optimized matrices, they do not reveal the extent to which any structural property enhances or detracts from transmission accuracy. We therefore perform correlational and regression statistics to expose relationships between matrix structure and transmission accuracy.

Several interesting trends emerge from statistical investigation of the distance values. We perform a step-wise linear regression, in which distance from the identity matrix, distance from the uniform matrix, and distance from the deterministic matrix are entered successively as predictors of transmission accuracy. Based on their Pearson correlations, transmission accuracy is positively correlated with distance from the identity matrix, $r = 0.368$, $p = 0.002$, but negatively correlated with distance from the deterministic matrix, $r = -0.223$, $p = 0.044$. Nonetheless, only distance from the identity matrix is a significant predictor of transmission accuracy when all three factors are entered into a regression model, $p = 0.004$. The resulting model has the following parameters:

Table 5.3: Parameters of Linear Regression Model of Transmission Accuracy

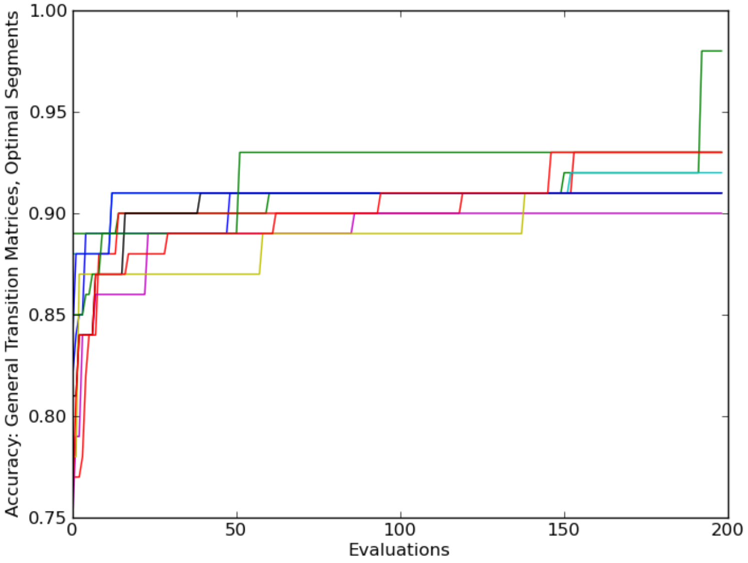
Variable	B	Standard Error	β
Constant	0.522	0.108	0.368
Distance from Identity Matrix	0.091	0.030	

As expected based on the Pearson correlations, transmission accuracy increases as a function of distance from the identity matrix, accounting for 36.8% of the variance in the system. Neither distance from a deterministic matrix nor distance from a uniform matrix contributes to a predictive model of transmission accuracy. The low R value for the regression indicates that although distance from the identity matrix is a significant predictor of transmission accuracy, there are other factors that account for the remaining 63.2% of the variance in the model. The relative unimportance of distance from a uniform or deterministic matrix indicates that the opportunity to include all dictionary segments is more important than the strategy for balancing those segments in message construction.

5.2 Condition II: Optimized Segment Dictionaries

Having optimized transition matrices over a standard random segment dictionary, we seek to determine whether optimizing a transition matrix over an optimized segment dictionary improves transmission accuracy. To that end, we select the segment dictionary that achieved the highest accuracy score in each noise condition and set it as the fixed dictionary for the corresponding condition in our evolving matrix simulation. As before, we perform ten runs of each optimization scenario, with run lengths depending on the number of evaluations required for the system to reach a stable maximum. A representative trajectory is presented below.

Figure 5.3: Evolution of Transmission Accuracy when Optimizing Transition Matrices Based on Optimized Segment Dictionaries



As before, all evolutionary trajectories have a positive slope; optimizing transition probabilities thus enhances communicative accuracy even when using an optimized dictionary. We compare the average accuracy values attained in the random and optimized segment dictionary conditions below.

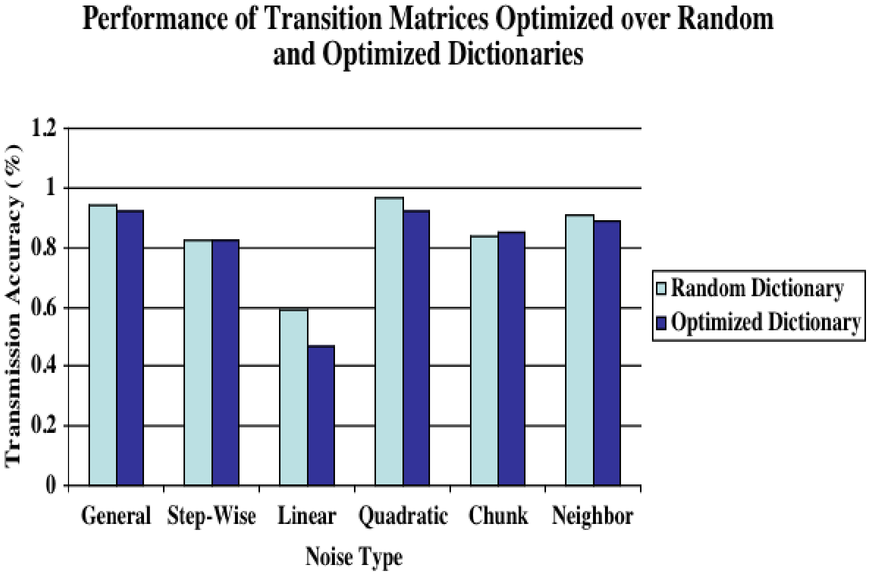


Figure 5.4: Transmission Accuracies of Transition Matrices Optimized Using Random and Optimal Dictionaries

The accuracies obtained using optimized segment dictionaries do not appear to differ substantively from those obtained with a random dictionary. In some cases, in fact, the optimized segment dictionary appears to have fared slightly worse than its random counterpart. We perform an independent samples t-test comparing the results of the two simulations; it is not significant, indicating that using optimized segment dictionaries does not substantially improve transmission accuracy.

5.2.1 Matrix Structure

Even if using an optimized segment dictionary does not affect the overall accuracy level attained in a given system, optimized dictionaries may exert different pressures on evolving transition matrices. We apply the same Euclidean distance analyses to the matrices evolved using optimized segment dictionaries to determine whether they differ structurally from those evolved using random dictionaries.

Table 5.4: Euclidean Distance of Optimized Transition Matrices from Three Canonical Matrices

Noise Type	Uniform Matrix	Identity Matrix	Deterministic Matrix
General	1.268	3.57	3.57
Std. Dev.	0.521	0.659	0.764
Step-Wise	1.348	3.628	3.569
Std. Dev.	0.27	0.596	0.476
Linear	1.03	3.469	3.498
Std. Dev.	0.329	0.59	0.52
Quadratic	1.22	3.523	3.758
Std. Dev.	0.225	0.586	0.377
Chunk	1.18	3.57	3.33
Std. Dev.	0.19	0.78	0.42
Neighbor	1.15	3.278	3.597
Std. Dev.	0.277	0.555	0.379

These distance values do not appear to differ significantly from those obtained using matrices evolved with a random dictionary, an assumption we once again assess using independent-samples t-tests. None of the comparisons is significant, permitting us to conclude that transition matrices optimized over random and optimized dictionaries are structurally equivalent.

We also perform a linear regression to investigate the predictive power of the three distance criteria in relation to transmission accuracy. In this condition, however, none of the distance variables is a significant factor affecting transmission accuracy, nor is the regression or corresponding ANOVA significant. Some significant relationships obtain among distance variables; as before, distance from the deterministic and identity matrices are negatively correlated, $r = -0.351$, $p = 0.003$. Unlike in the previous condition, however, distances from the uniform and deterministic matrices are significantly positively correlated, $r = 0.362$, $p = 0.002$.

Random and optimized segment dictionaries, then, produce functionally equivalent transition matrices; they achieve the same level of transmission accuracy and are structurally similar. Based on the results of the two linear regressions, however, random and optimized segment dictionaries produce transition matrices with slightly different properties. Although on average the matrix structure values are comparable across conditions, relationships among structural properties differ based on the refinement of the segment dictionary. Most striking is the disparity in predictive power; while increased structure predicts improved transmission accuracy in the random dictionary condition, structure does not affect transmission accuracy in the optimized dictionary condition. Coupled with differences by condition in relationships among structural properties, this contrast underlines the burden placed on transition matrices when using sub-optimal dictionaries. When dictionaries are not refined, transition matrices must become more structured in order to equal the transmission accuracy of matrices evolved with respect to optimized dictionaries. In essence, in the random condition the transition matrix bears the structural burden, but in the optimized condition the refined segment dictionary makes higher degrees of matrix structure unnecessary.

Here we confirm that statistical strategies enhance communication accuracy. As previously noted, the accuracy levels attained in both matrix optimization conditions are significantly higher than those achieved using an intentionally-constructed matrix. More important, however, is the role of transition matrix structure. In the absence of refined segment dictionaries, distance from the identity matrix – linguistically, greater structure – is a significant predictor of transmission accuracy. Although matrix structure is not a significant predictor of transmission accuracy in the optimized dictionary condition, natural selection converges on statistically equivalent matrices in both conditions, pointing to an underlying preference for more structured systems. We observe this structural convergence as a consequence of pressure favoring more accurate communication; in this case, then, linguistic structure is a by-product of increased transmission accuracy. We must therefore conclude that the efficiency of linguistic structure as an error-correcting code is not coincidental but rather represents the evolutionary importance of conveying information in 'noisy' environments.

Chapter 6

Non-Adjacency

The improvement in communicative accuracy consequent to employing probabilistic decoding strategies underlines the importance of likelihood relationships among segments in a linguistic system. Our previous simulation focused on likelihood relationships between consecutive segments, but natural languages also exhibit important structural relationships between non-consecutive units. If these long-range dependencies prove to be informationally adaptive, we support the theory that linguistic structure serves an explicitly code-like function.

In order to assess the communicative utility of non-adjacency relationships, we must develop digram and trigram transition matrices. Each row and column of each digram or trigram matrix represents a sequence of segments, and each entry represents the probability of transition from the row-sequence to the column-sequence. The entry in the row corresponding to the digram (1,0) and the column corresponding to the digram (0, 1), for example, gives the probability of observing segment 1 given that the previous set of segments was (1, 0). Mathematically, such a digram transition is written as a conditional probability: $P\{X_n = 1 \,|\, X_{n-2} = 1, X_{n-1} = 0\}$. Using a digram or trigram transition matrix allows us to observe the frequency with which segments recur at specified intervals.

A first impulse would be to repeat the previous matrix optimization procedure with a matrix representing digrams or trigrams, but we hypothesize that optimization may not be necessary to induce long-range dependencies. Non-adjacency relationships may emerge from frequency statistics accumulated over many messages constructed using the likelihoods encoded in a fixed inter-segment transition matrix. If so, the statistical algorithm is doubly effective as a coding strategy; it induces both immediate and long-term structure in the communicative system.

To test this hypothesis, we conduct two simulations, one in which we found our long-term transition probability assessment on an optimized segment dictionary and optimized inter-segment transition matrix, and a control condition in which both the segment dictionary and transition matrix are generated randomly. In both simulations we use the transition matrix and segment dictionary to compose a message that is four segments long, selecting the first segment at random from the dictionary and the remaining three stochastically based on the transition matrix, which we then expose to noise and decode. If the decoded message matches the original message, we record sequence transition frequencies and re-expose the decoded message to noise. For example, if we correctly decode a message composed of segments 0, 1, 1, 2, in that order, we increase the frequency with which the sequence (0, 1) is followed by the sequence (1, 1) and the frequency with which the sequence (1, 1) is followed by the sequence (1, 2). If the message is incorrectly decoded, we repeat the noise exposure and decoding procedure until we are successful. After 10000 iterations of noise exposure and decoding, we extract a transition matrix from the frequency dictionary and identify the most probable transition in each row of the matrix. Since each row of the matrix represents a possible sequence, these highly probable transitions should reflect long-range structural tendencies.

An apparent flaw in the design of this simulation is our failure to subject the digram and trigram transition matrices to pressure directly promoting transmission accuracy. Although we do not explicitly optimize the digram and trigram transition matrices, we only record transition frequencies from correctly decoded messages in order to reinforce communicatively adaptive structures. In so doing, we implicitly apply pressure favoring increased transmission accuracy.

We restrict message length to four segments in order to maintain a reasonable level of transmission accuracy. As noted in depth in previous chapters, some noise functions are more stringent than others, and an arbitrarily low success rate would compromise the amount of data on which the sequence transition matrices are based. To compensate for the small number of transitions present in any given message, we assess 10000 messages in each trial.

We conduct 25 trials of 10000 transmissions for each noise function at both the digram and trigram levels, first using an optimized segment dictionary and corresponding optimized transition matrix, then using a random dictionary and random matrix. Comparing these conditions should highlight the effect, if any, of refined communicative systems on long-range structure.

In both the random and optimized system conditions we observe convergence on a most probable segment on the digram and trigram levels. In every run of every noise condition, every sequence is most likely to transition to a sequence ending in the same segment. If the initial sequence is (0, 0), for example, the most likely next sequence is (0, 1); we are most likely to transition from (0, 1) to (1, 1), from (1, 0) to (0, 1), from (2, 0) to (0, 1) and so on. No matter what two-segment sequence we observe at the beginning of a three-segment sequence, the third segment is always most likely to be segment 1. The digram and trigram transition matrices in each noise condition favor transitions to one particular segment for every initial sequence.

These preferences, while interesting, represent only the most probable transition in a given row of each transition matrix. It is possible that the 'preferred' segment appears hardly more often, on average, than any of the other segments. Analyzing matrix structure using Euclidean distance criteria should highlight both the strength of these preferences as well as the amount of structure present in digram and trigram systems. Here we compute only the distance between each sequence transition matrix and an identity and uniformly distributed matrix of the same dimensions. As before, we regard these distances as proxies for the bounds on Shannon entropy.

As originally written, our distance equation computes the Euclidean distance between corresponding rows of two matrices and sums the row-distances to obtain the total distance between the matrices. A natural consequence of this method is that larger matrices have larger distances simply because they contain more rows. To facilitate comparison between our digram and trigram conditions, we scale each total distance by the number of rows in the matrix. We present the scaled average distances from identity and uniform matrices in tables 6.1 and 6.2.

Table 6.1: Euclidean Distance of Digram Transition Matrices from Uniform and Identity Matrices

	Optimized Systems		Random Systems	
Noise Type	Uniform	Identity	Uniform	Identity
General: mean	1.1	2.32	1.19	2.33
Std. Dev:	0.091	0.045	0.081	0.032
Step-wise: mean	1.36	2.43	1.13	2.298
Std. Dev:	0.038	0.052	0.08	0.026
Linear: mean	1.31	2.43	1.2	2.3
Std. Dev:	0.38	0.36	0.18	0.06
Quadratic: mean	1.23	2.27	1.27	2.34
Std. Dev:	0.14	0.042	0.11	0.042
Chunk: mean	1.36	2.33	1.02	2.22
Std. Dev:	0.07	0.086	0.046	0.041
Neighbor: mean	1.16	2.33	1.16	2.32
Std. Dev:	0.062	0.019	0.06	0.028

Table 6.2: Euclidean Distance of Trigram Transition Matrices from Uniform and Identity Matrices

	Optimized Systems		Random Systems	
Noise Type	Uniform	Identity	Uniform	Identity
General: mean	2.27	4.08	2.44	4.27
Std. Dev:	0.081	0.91	0.395	0.56
Step-wise: mean	2.49	4.29	2.416	4.27
Std. Dev:	0.44	0.36	0.68	0.44
Linear: mean	2.48	4.29	2.71	4.466
Std. Dev:	1.8	1.26	3.07	2.1
Quadratic: mean	2.51	4.18	3.09	4.73
Std. Dev:	2.00	0.983	1.12	0.824
Chunk: mean	2.467	4.21	2.67	4.45
Std. Dev:	0.53	0.417	1.33	0.872
Neighbor: mean	2.75	4.497	2.4	4.253
Std. Dev:	1.04	0.71	0.72	0.48

The scaled distances manifest clear disparities in magnitude. Across all conditions, the distance from the identity matrix is roughly twice the distance from the uniform matrix, throwing the systems' preference for maximal information into sharp relief. Second, trigram distances are roughly twice as large as digram distances. If we treat distance from the uniform matrix as a proxy for relative entropy, communicative systems defined by trigram transitions contain half the uncertainty – or twice the structure – of systems defined by digram transition matrices.

To determine whether the trend extends to segment transition matrices, we scale the relevant distance statistics appropriately and compare them to scaled digram and trigram distances. All trajectories demonstrate the same trend, so we select three representative trajectories and present them below.

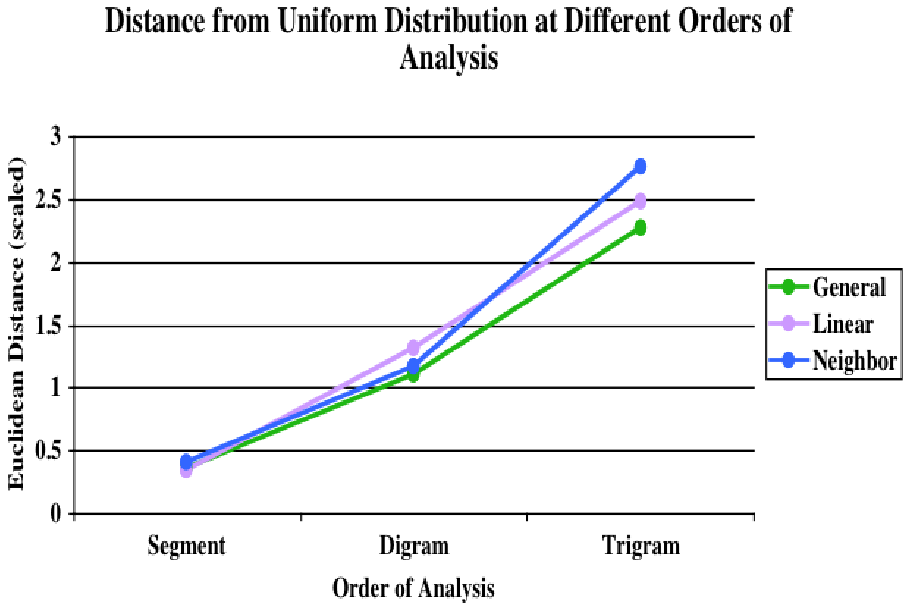


Figure 6.1: Increase in Structure with Increasing Orders of Analysis: Distance from Uniform Distribution

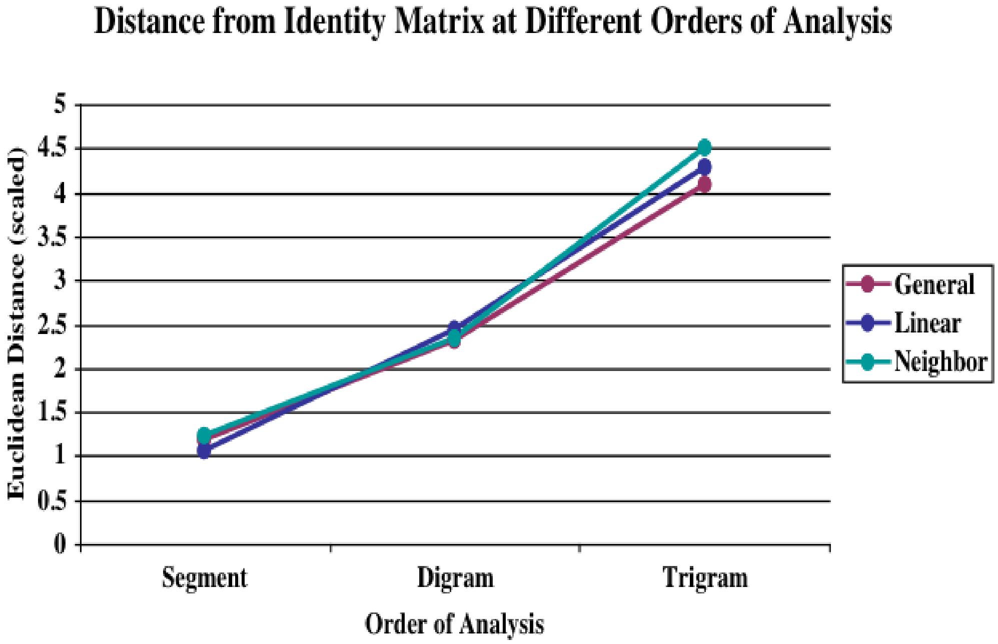


Figure 6.2: Increase in Structure with Increasing Orders of Analysis: Distance from Identity Matrix

The trend holds across all three orders of analysis; not only do distances from both the uniform and identity matrices increase monotonically, roughly doubling in each condition, but in all conditions the distance from the identity matrix is about twice the distance from the uniform matrix. If we treat distance from the uniform matrix as a rough measure of relative entropy, we observe a quasi-linear decrease in relative entropy with higher orders of analysis. At the same time, we observe a quasi-linear increase in distance from the identity matrix, suggesting that the non-adjacency relationships we note above reflect increasingly strong long-range dependencies. The simultaneous increase in distance from the uniform and identity matrices points to a balance between maximizing information and maximizing order.

For Shannon, this mid-point between complete uncertainty and complete determinacy represents the position of natural languages on an entropy continuum. He links a language's redundancy to the number of letters a native speaker can restore to a text from which they have been deleted (Shannon 1948: 25). Although he does not explicitly deduce that this property facilitates accurate decoding, it is evident from this example that a semi-redundant language – one in which linguistic gaps can be filled – is by definition easy to reconstruct.

We compare distance from the identity and uniform matrices in different noise conditions to determine whether different matrix structures emerge under different noise functions. The ANOVAs indicate that matrix structure differs significantly by noise condition; the relevant figures are presented in table 6.3.

Table 6.3: F and p-values for ANOVAs Comparing Euclidean Distances from Identity and Uniform Matrices by Noise Condition

Canonical Matrix	Sequence Length	Condition	F	p
Uniform	Digram	Optimized	953.78	1.9E^{-108}
		Random	333.06	5.98E^{-64}
	Trigram	Optimized	257.85	5.66E^{-70}
		Random	556.23	3.02E^{-92}
Identity	Digram	Optimized	632.1	4.7E^{-96}
		Random	251.83	2.59E^{-69}
	Trigram	Optimized	629.1	6.5E^{-96}
		Random	560.71	1.75E^{-92}

Post-hoc tests are also significant, $p < 0.001$ in all cases, indicating that each noise condition produces matrices that are structurally distinct from the matrices produced in all other conditions. No systematic trends emerge in these post-hoc tests, which suggests that sequence length, refinement of the initial communicative system, and noise type all play significant roles in shaping matrix structure; noise type alone is not responsible for the observed variation. Next we determine whether initiating the simulation with a random or optimized dictionary-matrix pair affects matrix structure. Random matrix structure differs highly significantly from optimized matrix structure for both digram and trigram transition matrices, all p-values $<< 0.0001$. The direction of the inequality, however, differs based on the order of analysis; in the digram case, optimized matrices exhibit greater distances from both uniform and identity matrices, but in the trigram case optimized matrices are closer to both uniform and identity matrices.

These statistics indicate that optimized matrix-dictionary combinations produce more structured digram systems. For trigrams, random dictionary-matrix combinations induce more structure. Although this conclusion appears counter-intuitive, it is important to recall that the underlying system only influences inter-segment transitions. Shannon himself observes that Markovian n-gram systems induce apparent structure to one level beyond that which they specify (Shannon 1948: 14), that is, generating strings using a matrix that defines conditional probabilities of transition between consecutive words produces sensible three-word strings, but nonsense strings of four words or more. In this case, we have defined the inter-segment transition matrix, so Shannon's rule of thumb predicts that the three-segment sequences will reflect the inherent structure of the inter-segment matrix and any longer sequences will diverge from those patterns. Since we use three-segment strings to record digram frequencies and four-segment strings to record trigram frequencies, it is likely, based on Shannon's observation, that the higher degree of structure observed in trigrams based on a random dictionary-matrix combination reflects the waning influence of the inter-segment transition matrix.

Having probed the differences between digram and trigram transition matrices extracted from random and optimized dictionary-matrix combinations, we wish to determine whether the structural effects we observe in these matrices reflect the selective pressure imposed by retaining only successfully decoded messages. To that end, we generate random matrices of the same dimensions and measure the Euclidean distance between those matrices and a uniform or identity matrix of the same dimensions. We scale these distances by the number of rows in the matrix and perform an independent-samples t-test to determine whether random and refined digram and trigram transition matrices differ structurally; the results of the t-tests are presented in table 6.4.

Table 6.4: Independent-Samples t-test Statistics Comparing Structural Metrics of Random and Extracted Digram and Trigram Matrices

Canonical Matrix	Sequence Length	t	p
Uniform	Digram	-108.06	$1.004E^{-103}$
	Trigram	-447.64	$4.72E^{-164}$
Identity	Digram	-16.687	$2.14E^{-30}$
	Trigram	-163.01	$3.99E^{-121}$

As is evident from table 6.4, the digram and trigram matrices extracted using our implicit optimization procedure demonstrate much higher degrees of structure than randomly-generated matrices of the same dimensions. We can therefore conclude that the increase in structure we observe with higher orders of analysis is due to the optimization procedure and not to inherent properties of matrices of those dimensions.

Long-range dependency relationships emerge from implicit optimization over digram and trigram transition matrices. Not only do significant patterns arise in digram and trigram transitions, but entropy decreases with higher orders of analysis. One can argue that these results are vacuous in the absence of explicit pressure for communicative success, but the retention of only successfully decoded messages as input for the digram and trigram systems acts as an indirect optimization. The only pressure present in this simulation favors transmission accuracy, and this pressure is demonstrably effective as a means of inducing structure; the highly significant structural disparity we observe between digram and trigram matrices subjected to this selective force and randomly-generated matrices of the same dimensions attests to the importance of pressure for transmission accuracy. We therefore conclude that the increased structure observed in higher-order representations of language emerges as a consequence of selecting for enhanced transmission accuracy.

Chapter 7

Conclusion

The existence of structural correspondences between language and information theory has long been accepted in linguistics. We contend that these structural resemblances are not superficial but rather reflect a causal connection between the two systems; linguistic structure, like code structure, conserves information under noise.

We observe that different types of error-correcting codes rely on varying degrees of structure, and select the Hamming code and the repetition code as archetypal strategies. Comparing the efficacy of the two codes at preserving information under noise, we identify redundancy and message length as the most important factors affecting transmission accuracy. Although the two codes reconstitute short messages equally accurately, the increased redundancy implicit in the repetition code results in higher accuracy for longer messages. Applying more complex noise functions reveals the superficiality of this apparent advantage; when message length is penalized, as in linguistic communication where longer messages introduce increasing opportunities for misinterpretation, the Hamming code is more accurate than the repetition code. In a purely information theoretic framework, internal structure is an optimal communicative strategy when the environment discourages unduly long messages.

We extend this logic to a linguistic-like message construction algorithm in which we restrict the set of possible bit sequences to three pre-determined four-bit segments, replicating the fixed phonetic or lexical inventories present in natural language. We expose messages composed in this way to a variety of noise functions and observe a statistically significant increase in transmission accuracy compared to the Hamming and repetition codes. We then introduce a transition matrix representing the varying probabilities of different segment sequences. This statistical method further improves transmission accuracy. Given the linguistic basis of these coding mechanisms, their success confirms the profundity of the similarity between linguistic and informational structure. The structural similarities between language and information theory are not merely topical; elements of linguistic structure, like fixed phonetic or lexical segment inventories and probabilistic relationships among those segments, enhance transmission accuracy. Linguistic structure preserves information under noise.

Although it is clear that linguistic structure is adaptive as a method of ensuring transmission accuracy, it is possible that structure is not specifically an adaptation for noisy communication – the correspondence could be a coincidence, a possibility we wish to eliminate. We therefore investigate the evolution of linguistic coding mechanisms under pressure for transmission accuracy to probe the nature of the relationship between noisy communication and linguistic structure.

We apply a hill-climbing technique to evolve communicatively optimal segment dictionaries of three, four, and six segments. Contrary to expectation, structural analysis of the resulting dictionaries reveals no systematic relationship between transmission accuracy and segment distinctiveness; instead, we remark a strong preference for homonymy. Transmission accuracy decreases with increasing dictionary size, pointing to a bias for homonymy over expressivity. Mutual distinctiveness, measured by maximum and minimum inter-segment distance, also declines with increasing dictionary size, but comparison to randomly-generated dictionaries indicates that this correlation arises from limitations on a finite combinatorial system.

Homonymy is thus the most prominent predictor of transmission accuracy in this paradigm, a result that flies in the face of recent research on the importance of maximal distinctiveness. Our apparently contradictory result, however, replicates the outcome of Kirby, Cornish, and Smith (2008), in which human participants manifested a strong preference for homonymous inventories when learning and transmitting an alien language in an iterated learning paradigm. The striking similarity of these two results strongly reaffirms the importance of transmission accuracy in the evolution of communication. More importantly, it establishes limitations on the influence of transmission accuracy on linguistic structure; although optimizing segment inventories with respect to transmission accuracy produces accurate systems, no structural trends emerge as a consequence of favoring transmission accuracy.

Evolving transition matrices using the same hill-climbing technique narrows the scope of the connection between transmission accuracy and higher degrees of linguistic structure. Optimized transition matrices based on random and optimized segment dictionaries attain higher levels of transmission accuracy than systems founded solely on optimized segment dictionaries, highlighting the centrality of probabilistic relationships to communication in noisy environments. Matrix structure, measured by distance from the identity matrix, is a significant predictor of transmission accuracy in the random dictionary condition, and matrices evolved using random and optimized dictionaries converge to statistically equivalent structural forms. Selective pressure favoring transmission accuracy induces probabilistic structure.

The linguistic structure observed in the optimized transition matrices reflects only relationships between consecutive segments. We then extract digram and trigram transition matrices from messages correctly decoded using optimized and randomly-generated dictionary-matrix combinations. Analysis of the structure of the resulting matrices reveals a striking trend; compared to inter-segment transition matrices, digram and trigram transition matrices exhibit higher degrees of internal structure as measured by Shannon entropy. Moreover, digram transition matrices extracted from optimized dictionary-matrix combinations manifest significantly higher structure than those extracted from random dictionary-matrix combinations, strengthening the importance of selective pressure in producing highly structured systems. Crucially, the increase in structure observed in digram and trigram transition matrices subjected to implicit pressure promoting transmission accuracy does not appear in randomly-generated matrices of the same dimensions, confirming our intuition that these structural trends are a direct consequence of pressure for transmission accuracy. Selecting for transmission accuracy produces more structured communicative systems.

The emergence of linguistic structure as a consequence of pressure on probabilistic relationships among linguistic units, but not in response to pressure on the units themselves, suggests that from an information theoretic perspective linguistic units are only beneficial insofar as they provide a stable framework for probabilistic interdependencies. Our simulations support the subordination of linguistic units to their relationships; as we observe in our optimization of inter-segment transition matrices, the refinement of the segment dictionary does not affect the transmission accuracy of the corresponding optimized transition matrix. Transmission accuracy is ambivalent with respect to segment structure, so other factors, like coordination among agents, are more likely to explain this aspect of linguistic communication.

Language's mathematical structure, however, is undeniably sensitive to pressure favoring transmission accuracy, throwing the centrality of mathematical properties underpinning language into sharp relief. Representing language as a complex probabilistic process grounds the comparison between linguistics and information theory, and our main result – the emergence of linguistic structure in response to selection for transmission accuracy – strengthens the connection between probabilistic form and linguistic function. It is only by reducing language to its mathematical components that we can observe and quantify the effect of transmission accuracy on linguistic structure, and it therefore becomes evident not only that language is explicitly mathematical but also that mathematical models are crucial to our understanding of language and its evolution. Here we have evaluated the evolutionary impact of modeling language as a simple discrete-time Markov chain, but more sophisticated models involving nested or non-homogeneous stochastic processes are likely to offer both more accurate approximations of natural language and glimpses of its evolutionary origins. Exploring the relationships between mathematics and language provides vital insight into its structure and opens a fruitful avenue of inquiry into the forces shaping its evolution.

Works Cited

Aslin, Richard N., Saffran, Jenny R., and Elissa L. Newport, 1998. "Computation of Conditional Probability Statistics by 8-Month Old Infants." *Psychological Science* 9 (4): 321 - 324.

Cherry, E. Colin, 1951. "A History of the Theory of Information." *Proceedings of the IEE - Part III: Radio and Communication Engineering* 98 (55): 383 - 393.

de Boer, Bart, 2000. "Emergence of vowel systems through self-organisation." *AI Communications* 13: 27 - 39.

Harris, Zellig, 1991. *A Theory of Language and Information.* Oxford: Oxford University Press.

Kirby, Simon, 2001. "Spontaneous Evolution of Linguistic Structure – An Iterated Learning Model of the Emergence of Regularity and Irregularity." *IEE Transactions on Evolutionary Computation* 5 (2): 102 - 110.

Kirby, Simon, Cornish, Hannah, and Kenny Smith, 2008. "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language." *Proceedings of the National Academy of Sciences USA* 105 (31): 10681 - 10686.

Liljencrants, Johan, and Björn Lindblom, 1972. "Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast." *Language* 48 (4): 839 - 862.

Plotkin, Joshua B., and Martin A. Nowak, 2000. "Language Evolution and Information Theory." *Journal of Theoretical Biology* 205: 147-159.

Saffran, Jenny R., Aslin, Richard N., and Elissa L. Newport, 1996. "Statistical Learning by 8-Month-Old Infants." *Science* 274 (5294): 1926 - 1928.

Shannon, C.E., 1948. "A Mathematical Theory of Communication." In Sloane, N.J.A. and Aaron D. Wyner (eds), *Claude Elwood Shannon: Collected Papers.* New Jersey:

Institute of Electrical and Electronics Engineers, 1993. pp. 5-83.

—, 1950a. “The Lattice Theory of Information.” In Sloane, N.J.A. and Aaron D. Wyner (eds), *Claude Elwood Shannon: Collected Papers.* New Jersey: Institute of Electrical and Electronics Engineers, 1993. pp. 180-183.

—, 1950b. “Prediction and Entropy of Printed English.” In Sloane, N.J.A. and Aaron D. Wyner (eds), *Claude Elwood Shannon: Collected Papers.* New Jersey: Institute of Electrical and Electronics Engineers, 1993. pp. 194-208.

Tervo, Richard, 2005. “Error Correction and the Hamming Code.” University of New Brunswick. www.ee.unb.ca/tervo/_ee4253/hamming.shtml. Viewed 9 August 2011.

Zuidema, Willem, and Bart de Boer, 2009. “The evolution of combinatorial phonology.” *Journal of Phonetics* 37 (2): 125-144.

— and Gert Westermann, 2003. “Evolution of an Optimal Lexicon under Constraints from Embodiment.” *Artificial Life* 9 (4): 387 - 402.