Information:
A theory of human communication

Michael Ramscar

Introduction

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design."
>
> C. E. Shannon, (1948), "*A Mathematical Theory of Communication*" p.31

> "the hard core of information theory is, essentially, a branch of mathematics, a strictly deductive system. A thorough understanding of the mathematical foundation and its communication application is surely a prerequisite to other applications. I personally believe that many of the concepts of information theory will prove useful in these other fields-and, indeed, some results are already quite promising-but the establishing of such applications is not a trivial matter of translating words to a new domain, but rather the slow tedious process of hypothesis and experimental verification. If, for example, the human being acts in some situations like an ideal decoder, this is an experimental and not a mathematical fact, and as such must be tested under a wide variety of experimental situations. "
>
> C. E. Shannon, (1956) "*The Bandwagon"*p3

How is it that human beings are able to communicate meaning through the medium of language? Why, as you read these typed words, do you form thoughts that approximate the ones I envisaged you having as I sat on my couch typing them? What, precisely, goes on as we communicate with words, and what makes it possible?

In what follows, I will briefly describe what I will term the “standard” conceptual model of linguistic communication, and detail several more or less well known shortcomings from which it suffers. I will suggest that these can be distilled down to a single essential point: that the standard model is non-computational, in that it involves making assumptions that violate some very basic computational and mathematical principles of coding. Accordingly, I will suggest that if we are to assume that our minds are best understood computationally (and I will describe some of the reasons why I believe that this is so), it follows that there must be a more profitable way of conceiving of language than the standard model.

I will then describe an alternative model of human communication grounded in formal theories of information and learning, and describe some of the evidence that supports this alternative model. I will argue that the alternative model of language I present not only offers a far more profitable and productive avenue for the scientific study of language, but also that it may have benefits for the users of language as well, because improving our understanding of the ways and means of the practices of human communication can allow those practices to be refined and improved on.

What is language? The traditional answer

Traditionally language has been seen as a system for communication that comprises a taxonomy of signs and meanings, and a code for combining simple signs into complex signs. Although signs can be composed of sounds, gestures, letters or symbols (or even smaller sub-symbols) depending on whether the language is spoken, signed or written, the assumption that unifies traditional approaches to language is that a relatively fixed vocabulary of signs can be combined into larger complex signs such as words and phrases by the use of rules. These rules are often thought of as comprising the syntax, or grammar, of a language, while the meanings that are connected to individual signs, words and phrases are called it semantics.

For the present purposes, the most important conceptions that the standard model embraces are the idea that words refer in specific ways to their meanings (and, often, things in the world), that their meanings can be combined, and that when a sign is used in communication, ands that meaning can be encoded and transmitted by a sender through a channel to a receiver who decodes it (a signal). That is, on the standard model, it is assumed (albeit often implicitly) that the information in the transmission channel (be it spoken, signed or written) is sufficient for the *meaning* of a message to be decoded from it.

Computationally, in order for meaning to be encoded into and decoded from a given token of a sign, type/token relationships for classed of signs to be defined. Defining what constitutes an X or a Y enables individual X's and Y's to be bound to the appropriate part of a structure such as "if X then Y," allowing the structure to describe a relationship in the world, and for the meaning of X and Y to be decoded from a message in which they are mentioned. If the definitions of classes are themselves signs, this in turn imposes a requirement that any signs in the definitions be defined as well (i.e., if the definition of X is "all X's have Z," one needs to define Z).

The computational problem this approach runs into is that the "Russian Doll' strategy of defining signs in terms of further signs is inherently regressive. The sign "dog," is a token of the type "noun," just as "spaniel" is a token of "dog," and "Fido" is a token of "spaniel." In order to explain (or generalize) the relationships between "Fido," "dog" and "spaniel," it is not sufficient to say, "Fido is a spaniel" and "A spaniel is a dog." One must say *why* Fido is a spaniel, and why spaniels are dogs (as opposed to anything else). Similarly, saying, "a sentence is grammatical if it is syntactically correct," explains little unless one defines which things in the world that are and are not members of the classes "sentence," "grammatical" and "syntactically correct."

Further, if signs are conceived of as being "compositional"—such that complex signs such as sentences in natural language have meanings that are determined both the structure of sentence and the specific meanings of the signs out of which they are composed (see e.g., Fodor, 1998)—one needs an account of how one can somehow extract the relevant individual tokens of meaning from descriptions that only mention types; for example, one needs to be able to say *which* aspects of the meanings of "cat" "sat" and "mat" are relevant to the meaning of "the cat sat on the mat."

Neither desideratum is satisfied by any existing approach to semantics (Fodor, 1998; Murphy, 2002). Indeed, there are good reasons to believe that they cannot be satisfied: the kinds of things that people represent and think about symbolically do not fall into discrete classes of X's, Y's or Z's. Symbolic categories do not possess discrete boundaries (i.e., there *are no* fixed criteria for establishing whether an entity is an X or a Y) and entities are often assigned to multiple symbolic classes (i.e., they are sometimes X's; sometimes Ys). As a result of these and many other factors, symbolic type/token relationships appear to be inherently underdetermined (see e.g., Wittgenstein, 1953; Quine, 1960; Fodor, 1998). This is a serious problem for all current symbolic approaches (Fodor, 1998), and has prompted theorists to conclude that while there *must be* a solution, it is innate and largely inscrutable (Fodor, 1983; 1998; Chomsky, 2000):

> "acquisition of lexical items poses Plato's problem in a very sharp form. As anyone who has tried to construct a dictionary or to work in descriptive semantics is aware, it is a very difficult matter to describe the meaning of a word, and such meanings have great intricacy, and involve the most remarkable assumptions, even in the case of very simple concepts, such as what counts as a possible "thing." At peak periods of language of language acquisition, children are "learning" many words a day, meaning that they are in effect learning words on a single exposure. This can only mean that the concepts are already available, with all or much of their intricacy and structure predetermined, and the child's task is to assign labels to concepts, as might be done with very simple evidence."
>
> Chomsky (1997 p. 28-29)

Many of these problems arise because signs are traditionally conceived of in *referential* terms (Ramscar, Yarlett, Dye, Denny & Thorpe, 2010). The standard theory assumes that signs both represent and point to meanings, so that signs and their meanings share a *bi-directional* relationship (which allows for the encoding and decoding of meaning). Signs are seen as abstract representations that either *exemplify* (stand for) or *refer* (point) to their meanings (*referents*). These meanings can in turn be defined by reference to things in the world—that is, signs can be defined by reference to objects and events they refer to. So, for example, the sign "dog" may be defined by reference to a class of things in the world, *dogs*. The problems with this approach are largely the same as for type/token definitions, and have been laid out exhaustively (see e.g., Wittgenstein, 1953; Quine, 1960; Fodor, 1998; Murphy, 2002).

The assumption that signs and meanings enjoy a bi-directional relationship is at odds with the idea that symbols are *abstract* representations, because abstraction is not a bi-directional process. Abstraction involves reducing the information content of a representation, such that only information relevant to a particular purpose is retained (Hume, 1740; Rosch, 1978). In Shannon's terms, abstraction is not lossless, and as such, cannot be a bi-directional process: one can abstract *from* a larger body of information *to* an abstract representation of it (such as reading an article and summarizing it in an abstract), but one cannot reverse the process. The idea of "reverse abstraction" makes little computational sense because information discarded as part of the process of abstraction cannot be recovered from the abstraction (just as one cannot get the detailed methods and analysis sections from the abstract of a research article that one has never seen; see Ramscar, et al 2010).

Thus, in terms of the semantics of signs, the standard model comes with two large problems. First, the taxonomy of signs and meanings it presupposes must somehow be inductively gotten from the evidence facing a learner or theorist (or, if we are to adopt a nativist approach, one must explain how every possible concept a child might need to lexicalize has come to be encoded in the genome), and second, even allowing for that, one must explain how one can get the meaning of a sign from a token that abstractly "represents" that meaning.

Finally, of course, as is well known, even ignoring the problem of decoding the semantics of signs, the induction of the rules that are supposed to combine simple signs into complex signs in the standard model (i.e., syntax) also poses problems for the learner and the theorist. As far as the learner goes, many theorists have argued that the task of learning the rules is impossible, and that some knowledge of linguistic structure must be innate.

To summarize then, the standard model of language has led theorists to suppose not only that children are born with knowledge of every possible lexical concept that they might ever learn, but also knowledge of the syntax by which these concepts will be combined (Fodor, 2000; Chomsky, 2000). Indeed, Yang (2006) suggests that children are born knowing the grammar of every possible language, and that they face the task of unlearning all but the correct one. It is worth noting at this point that the difficulties that accounting for the semantics of statements relating to counterfactuals[1] in the standard model has led philosophers to argue that all and every possible world must exist (Lewis, 1986). While it seems that no-one has actually proposed that children are born with innate knowledge of every possible world as well as every possible meaning (yet, at least), this may be that no one has yet thought to explain how children ever come to understand the semantics of counterfactuals from this perspective. It would appear then that by adopting the standard model, researchers have suffered a series of theoretical and empirical defeats (in that study of most of the linguistically and psychologically interesting aspects of language has been put on hold until such time as advances in the study of genetics makes further progress possible). What's worse, however, it that for many, it appears that the scientific war has been lost as well: much research on the standard model now makes a distinction between people's idealized linguistic competence, and their actual linguistics performance, which can be summarized, briefly, as follows: language is possible only because of rules that are not evident in people's use of language, and thus *competence*, the linguistics rules that make language possible, can and should be studied independently of the only evidence for them (people's actual use of languages, *performance*; Chomsky, 1965); for many, the scientific study of language has been replaced by a pre-occupation with the philosophy of science (Nunberg, 1996).

What is language? An information theoretic approach

Historically, theoretical approaches to human communication have overwhelmingly adopted a generative, taxonomic approach to explaining how semantic information is communicated in text or speech. Linguistic communication has been seen as a process of encoding, transmitting and decoding tokens of meaning types. These meaning types have been assumed to be taxonomically organized, and encoded and decoded by rules that allow messages to be generated from them. From this perspective, the challenge facing both language learners and theoretical linguists is inductive: the correct taxonomy of meaning types and generative rules for a given language must be inferred from whatever data is available to the learner of theorist. As I have tried to show in

[1] Statements that run counter to what actually happens in the world, such as, "If JFK had not been shot…"

the brief review above, the evidence of history suggests that the problems these inductive challenges pose are insoluble.

However, in contrast to the inductive models that have dominated linguistics (and the psychology of language), both information theory (and the study of artificial communication systems) and empirically grounded psychological theories of learning describe deductive processes based on prediction and discrimination (Shannon, 1948; Weaver & Shannon, 1963; Kullback & Leibler,1951; Rescorla & Wagner, 1972; Ramscar et al, 2010). Seen from this perspective, communication need not require the induction of taxonomic semantic classes and rules for their combination; nor need it involve the transmission of tokens corresponding to semantic types between speakers.

To explain why, it is worth describing in some of the important empirical and theoretical work in information theory and learning theory that I shall suggest, taken together, can provide a far more productive alternative to the standard model of language.

Discrimination, information and “The Bandwagon”

To help understand the importance role discrimination and prediction play in information theory, it is worth reflecting on how Shannon’s seminal (1948) results arose out of earlier work on communication by Hartley (1928) and Nyquist (1924). While Nyquist’s (1924) focus was firmly on telegraph engineering, the problem he addressed – how to communicate between two points while eliminating errors that are likely to be introduced by the communication channel – is fundamental to information theory, and his paper contains two important results that influenced its development. The first of these was a logarithmic rule to describe the maximum amount of “intelligence” that could be sent across a telegraph wire:

$$W = K \, log \, m \tag{1}$$

where W is the speed of transmission of “intelligence”, m is the number of current values, and K is a constant. This law is not sufficiently general to describe all communication systems, however it is a special case of Shannon’s later logarithmic law, and importantly, it can be seen as representing the beginning of a move to abstract the message contained in a signal away from its actual meaning. In addition, Nyquist showed how the message content of a signal could be encoded “optimally” in order to permit the transmission of the maximum amount of “intelligence” with any given number of signal elements.

Hartley (1928) generalized many of Nyquist’s ideas, and was the first to formalize the differentiation between the information in a particular message and its meaning. In estimating the capacity of the physical system to transmit information, he showed, the question of interpretation could be ignored. Hartley demonstrated that each aspect of a signal could be considered as a perfectly arbitrary choice, and capacity could be determined by measuring the probability of a receiver distinguishing a given signal that was selected from that of selecting any of the other signals that might have been chosen. In doing this, Hartley generalized Nyquist’s logarithm law for the amount of information transmitted to:

$$H = log\, S^{n} \tag{2}$$

where *S* in the number of possible symbols in a code, and *n* is the number of symbols in a transmission.

The insights gained from this earlier work are immediately apparent in Shannon’s (1948) statement of the problem his theory of information sought to address:

> “The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.”

Shannon's paper presents a number of fundamental results: the first of these, the source coding theorem for symbol codes establishes that, on average, the number of bits needed to represent the result of an uncertain event is given by its entropy (this described in more detail below). The noisy-channel coding theorem then states that reliable communication is possible over noisy channels provided that the rate of communication is below a certain threshold, called the channel capacity, and finally the paper shows how using appropriate encoding and decoding systems allows this to be approached in practice.

Importantly, Shannon solved the problem of communication over a noisy channel (McKay, 2003) by focusing on the "system" itself rather than the channel (this approach was also advocated by Hartley). The system solution accepts the noisy channel as a given and focuses on the use of a communication system to overcome the limitations imposed by the channel by detecting and correcting the errors it introduces.

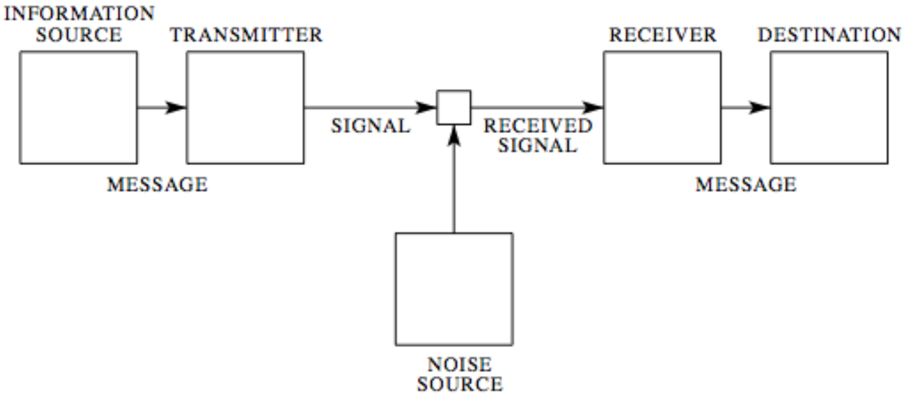

Figure 1. A schematic diagram of a communication system (from Shannon, 1948)

The system envisaged by Shannon comprises an *information source*, and a *destination*—as one might expect to find in any communication system—and adds an *encoder* before the *communication channel* and a *decoder* after it. The encoder encodes the source messages into a transmitted message, which adds redundancy to the original message. The channel then adds noise to the transmitted message, yielding a received message that comprises a mixture of the source message and the noise. The decoder then uses the known redundancy introduced by the encoding system (and the coding system itself) to discriminate the original signal from both the added noise and other possible signals.

Shannon's results were an enormously significant contribution to the development of modern communications and computer systems (if you are reading this on an electronic device, then the design of the technology that allows for both the storage of these words and their successful transmission to the interface on which you are reading them can trace its lineage directly back to Shannon's original theorems), and they were immediately received with much interest outside of communication engineering. Applications of information theory were soon being made to biology, psychology, linguistics, physics, economics, the theory of organization, and many other fields, prompting Shannon (1956, in an article entitled, *The Bandwagon*) to warn:

> "Information theory has, in the last few years, become something of a scientific bandwagon. Starting as a technical tool for the communication engineer, it has received an extraordinary amount of publicity in the popular as well as the scientific press. In part, this has been due to connections with such fashionable fields as computing machines, cybernetics, and automation; and in part, to the novelty of its subject matter. As a consequence, it has perhaps been ballooned to an importance beyond its actual accomplishments… workers in other fields should realize that the basic results of the subject are aimed in a very specific direction, a direction that is not necessarily relevant to such fields as psychology, economics, and other social sciences. Indeed, the hard core of information theory is, essentially, a branch of mathematics, a strictly deductive system. A thorough understanding of the mathematical foundation and its communication application is surely a prerequisite to other applications.. "

Although I will not try to elucidate the mathematical foundations of information theory here (not least because others are far better capable of doing so, see e.g., McKay, 2003), I will try to underscore the logic of its application to communication, because it is likely that it was inattention to this logic that prompted the bubble of enthusiasm for information theory in psychology and linguistics to which Shannon alludes, and also the subsequent deflation of that bubble.

First and foremost, it is important to understand what Shannon's system solution does and does not require, and what it does and does not involve:

1. The system solution requires a source encoder and a decoder that are both equally well aware of the scope of the possible messages that can be transmitted across the channel.

2. The system solution is not at all concerned with the *meaning* of messages. The goal of Shannon's system solution is that the receiver be able to successfully *reconstruct* the source message from the received message by discriminating the source message from other possible messages that might have been selected, and from noise introduced been by the communication channel.

3. The purpose of the decoder is not to interpret or expand on the source message in any way. It is simply to reproduce the source message at the destination with no loss of signal content.

I hope that the difference between communication as envisaged in information theory and communication as envisaged by the standard linguistic model is very apparent from these points. Shannon's system solution is designed to allow the receiver to guess the message encoded at the source with a high degree of accuracy. The use of the word "decode" in information theory does not correspond to the idea of a listener decoding a speaker's meaning in the standard model of language; it merely corresponds to the listener being able to successfully select the words a speaker did use from those a speaker might have used.

It follows from thus that any direct application of information theory to questions in psychology or linguistics requires a very different model of language to the one normally applied in those domains. If one's model of language involves the sending of messages and the extraction of meaning from them (and as I suggested above, explaining how this latter part happens is the biggest problem facing the standard model), then information theory may well be largely irrelevant to one's concerns, since its application requires a model in which the channel aspects of linguistic communication involve no more than a listener being able to successfully reconstruct whatever words a speaker utters. It is clear that Shannon (1956) was very aware of these important differences:

> "I personally believe that many of the concepts of information theory will prove useful in these other fields-and, indeed, some results are already quite promising-but the establishing of such applications is not a trivial matter of translating words to a new domain, but rather the slow tedious process of hypothesis and experimental verification. If, for example, the human being acts in some situations like an ideal decoder, this is an experimental and not a mathematical fact, and as such must be tested

> under a wide variety of experimental situations."

It is to Shannon's last question that I will now turn my attention.

Learning, discrimination and the surprising capacities of rats

> "So the question is, under exactly what circumstances does a child conclude that a nonwitnessed [form] is ungrammatical? This is virtually a restatement of the original learning problem. Answering it requires specifying some detailed learning strategy. It takes the burden of explaining learning out of the environmental input and puts it back in the child. Use of indirect negative evidence… is thus, not strictly speaking, a feature of the child's learning environment… but rather, a feature of his learning strategy, and hence it must be fleshed out according to a particular theory of these learning strategies." (Pinker, 1989)

In his 1989 book, *Learnability and Cognition*, Steven Pinker considers – and rejects – the possibility that 'indirect' negative evidence could provide a solution to problems facing a child learning language over the course of a single page. Pinker's approach is not unusual; it simply reflects a set of beliefs that have come to dominate our understanding of children's language learning over the past half-century. The standard model of language assumes that if children can use language, they must have mastered abilities that no amount of experience would appear to be able to equip them with. Accordingly, as I noted earlier, many proponents of the standard model of language have claimed that if the mechanisms the model supposes don't appear to be learned, they must be innate. As Pinker's remarks illustrate, this has in turn led many proponents of the standard model of language to actively argue *against* the possibility that language might be learned.

To understand the rather bemusing consequences of this approach, one has to turn one's eyes away from the realm of language learning and look to the more humble world of the laboratory rat: because for the past forty years, psychologists studying animal behavior have accepted that explaining the behavior of rodents requires fleshed out theories of learning that endow their furry subjects with abilities that go far beyond anything a language researcher would suppose a child might do. A large body of evidence supports that idea that rats' expectations provide a critical source of evidence across a wide range of learning tasks, and psychologists studying rats and mice have found that it is impossible to explain the cognitive behaviors of their rodent subjects unless they acknowledge that they are capable of learning in ways that are far more subtle and sophisticated than many researchers studying language would ever countenance in children (Rescorla-Wagner, 1972; Dyan & Daw, 2008; Ramscar et al, 2010).

Not only have animal learning models been fleshed out in ways that embrace the idea that animals make extensive use of indirect evidence in learning, but the computational properties of these models have been extensively explored (see Danks, 2003 for a review), and the biological circuits that implement these mechanisms are fairly well understood (Montague et al., 1996; Schultz, 1997; Waelti, Dickinson & Schultz, 2001; Schultz, 2006; Niv, 2009). More, as I will endeavor to show, research on animals has shown that while, say, rats may not have been blessed by nature when it comes to communicative capacities, those abilities that they do have would appear to successfully approximate the qualities one would expect to find in an ideal Shannon decoder (Gallistel, 2003).

As with information theory, a little history may help in explaining this point, becaused while much of our contemporary understanding of animal learning has its origins in Ivan Pavlov's (1927) classical conditioning experiments, it is critical to note that the conception of learning that people usually draw from Pavlov's work is, in most ways, the exact *opposite* of the understanding produced by the revolution that has occurred in the field of animal learning in the century following Pavlov's initial discoveries (Rescorla, 1988).

As is well known, Pavlov discovered that if he rang a bell as he presented food to a group of dogs, they soon began to salivate on hearing the bell, even when no food was on offer. These findings gave rise to the view that Pavlovian learning is a process of **association**, in which animals learn to 'associate' previously unrelated things in the world, such as a bell and a meal, by tracking the degree to which a **stimulus** (a bell) and a **response** (salivation brought on by food) are paired. Unfortunately, this also led to the mistaken view that animal learning is a simple process measuring the co-occurrence of events. This (false) idea, that learning resulted automatically from the repeated co-occurrence of a stimulus and a reward or punishment, in turn gave rise to two popular misconceptions about the conditions that are necessary and sufficient for learning: first, that explicit 'rewards' or 'punishments' are necessary for learning (false); and second, that a simple co-occurrence between a "stimulus" and a "response" is sufficient for learning: i.e., if a bell is simply paired with food often enough, a dog will always learn the association (wrong again; Rescorla, 1988).

A single experiment can show how both of these ideas are wrong (Rescorla, 1968): consider a variant of the classic Pavlovian conditioning paradigm in which rats are conditioned to associate a tone with mild electric shocks. The schedule of tones and shocks given to a first group of rats might look like this:

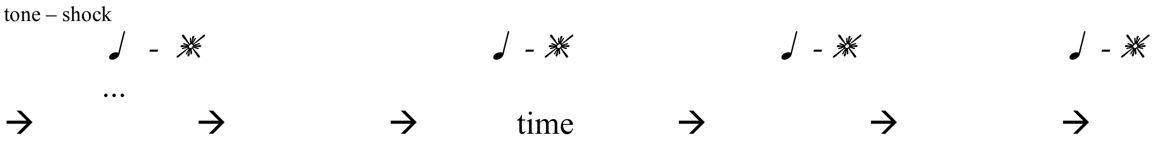

Figure 2. Schematic of a training schedule used by Rescorla (1968). Here, the rate of tones absent shocks is 0.

Like Pavlov's dogs, rats trained in this way will quickly learn to 'associate' the tones with the shocks, freezing whenever a tone sounded. If we now take another group of rats and expose them to the same number of tones followed by shocks as the first group, while also introducing a number of tones that are not followed by shocks (Rescorla, 1968):

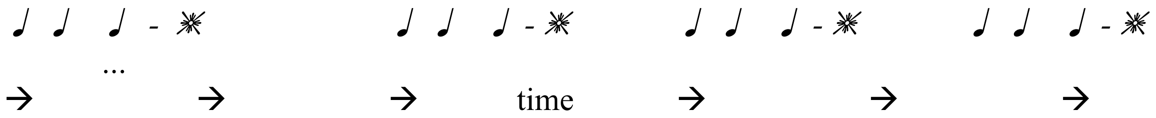

Figure 3. A training schedule in which the background of tones absent-shocks increases. Although the absolute number of tones leading to shocks is identical, only 1 in 3 of the tones is now followed by a shock, and the degree to which rats condition to the relationship between the tones and the shocks diminishes proportionally (Rescorla, 1968).

As the number of tones that don't result in shocks increases, rats will come to associate the tones with the shocks less and less. Indeed, if the **background rate** of the tones is varied (i.e., if the number of tones that lead to nothing is increased and decreased), the degree to which a rat will freeze upon hearing the tones will decrease in direct proportion to the background rate (Rescorla, 1988). Indeed, as the background rate of tones increases, conditioning decreased until the rats' responses to the tones all but disappeared, even though the rate at which the tones co-occurred with the shocks remained exactly the same.

This has important implications for the naïve model of learning I described above (Recorla, 1988): Given that there is no change in the 'association rate' between A and B – only the background rate varies – it follows that the difference in learning between the groups of rats is due to what the rats learned on the "no shock" trials. It follows, therefore, that if the differences in what the two groups of rats learns is driven by the *non*-occurrence of expected events, learning cannot simply be a process of tracking the co-occurrences between cues and events.

There is, however, more to learning than simply counting successful and unsuccessful predictions; because learning serves to increase the reliability of future predictions, it is also competitive. **Blocking** (Kamin, 1969) is a simple statistical consequence of the way learning serves to reduce uncertainty about future events: once a learner has learned to fully predict an outcome from a given cue, learning about additional cues is unnecessary, becuase the information provided by them will be redundant. Informally, one might think of uncertainty as a finite resource, which is depleted as successful predictive cues are learned. When a set of cues already fully predicts an event, there will be little uncertainty available to drive the learning of other cues to that event. As a result, cues can be seen as competing with one another for predictive value.

Blocking is easily illustrated in regards to animal learning: for example, if a rat has already learned that it will be shocked when it hears a tone, and a light is subsequently paired with the tone in training, the rat will fail to learn to value the light as an additional predictive cue. Because the tone is already fully informative about the upcoming shock, the information provided by the light is redundant, and so the rat's prior learning about the tone **blocks** subsequent learning about the light. As results like this demonstrate, rats do not learn simple "associations" between stimuli and responses; rather, they learn the degree to which cues are systematically *informative* about their environments.

This idea – that rats learn to predict, and in some sense ‘understand’ the world around them by monitoring when their predictions are successful and when they result in errors – can be formalized mathematically in a relatively straightforward manner (Rescorla-Wagner, 1972; Mackintosh, 1975; Pearce & Hall, 1980; Dickinson, 1980; further, the error-monitoring captured by formal models of animal learning appears to have a straightforward neural implementation: Schultz, Dayan & Montague, 1997; Schultz & Dickinson, 2000; Waelti, Dickinson & Schultz, 2001; Daw & Shohamy, 2004; Schultz, 2006, 2010).

Given the logic of this discussion of error and expectation, it may seem natural to wonder – *but what expectations? Which errors?* For instance, given that the rats in Rescorla’s experiment had no a priori knowledge about the relationship between the tones and the shocks, it should strike us as suspect that only the background rate of the tones seems to matter in their learning. This is because, of course, it *isn’t* the only thing that matters. In principle, everything in the rat’s experience and environment matters in predicting the shock (Rescorla, 1988). However, in the same way that increasing the background rate of the tones decreased their relevance in predicting shocks, the rat’s previous experience with other aspects of its environment has resulted in learning about their background rates, and this prior learning influences – and indeed, is integral to – the rat’s subsequent learning. For example, the degree to which the color of a rat’s enclosure is learned as cue to the shocks, will be affected both by the rate at which the color of the enclosure has been previously experienced in situations that did not lead to shocks, and the rate at which the color of the enclosure is subsequently experienced in situations that do not lead to shocks. Thus, in principle, learning about the tones takes place against the backdrop of the rat’s entire experience with every aspect of its environment. Learning, then, can be seen as a process that samples cues in the environment, while subjecting them to a competition process that biases what gets learned – discriminating in favor of more informative cues to events, and discriminating against less informative cues (Rescorla, 1988).

For the sake of simplicity, models and explanations will tend to focus on the most novel – and hence, most informative – cues in a given environment, while ignoring other potential cues, whose background rates are likely to be so high as to render them irrelevant in competitive terms. However, it is important to understand that the novelty of the tone is entirely relative: it only makes sense to assume that the tone is novel if we also assume that every other available perceptual cue is *not* as novel. Accordingly, in relation to a rat learning from its environment, it only makes sense to describe the tone as a 'novel' perceptual cue if we assume that this novelty has been computed in relation to all of the other perceptual cues available to the rat (Rescorla, 1988; Ramscar et al, 2010; this point may help clarify why animal researchers formally relate learning to a "stimulus complex," Rescorla & Wagner, 1972, rather than an individual stimulus).

Finally, it should be emphasized that in discrimination learning, the value of positively informative relationships – such as the predictive value tones might have in relations to shocks – is learned as a result of the competitive elimination of less informative relationships. Since the latter inevitably outnumbers the former, it follows, perhaps surprisingly, that expectations that are *wrong* have more influence on the shape of learning than expectations that are *right*.

These last points underline the relationship between learning and information. Since the late 1960s, "associative" learning has actually been understood as a discriminative process driven by prediction. What this means is that the "goals" of animal learning systems might be re-described, in Shannon's terms, as a process of decoding information available in the environment.

Consistent with this, Gallistel (2003) has shown that if we assume that the messages to be decoded are not sent from a source, but rather provided by the environment, animal learning can be seen as approximating an ideal decoder. Learned can be formalized in terms of information provided by events in the environment, and the informativity of these events is shaped by what the animal already knows, because any prior learning that will determine how much "information" is any new experience, and in what way it is informative to a learner. Gallistel (2003) shows how by adopting this perspective, the mechanisms that have been proposed to account for phenomena associated with animal learning (e.g., Rescorla & Wagner, 1972) can be shown to be formally equivalent to an information theoretic model in which learning is driven both by the information in the environment (the source message), and what an animal has learned as the result of previous experience (the source code). In Gallistel's (2003) model, learning only happens in response to informative events, and it can only occur if the amount of information in an event does not exceed a given threshold (the channel capacity; see also Gallistel & Gibbon, 2002). From this perspective, paralleling Shannon's definition of communication, the fundamental problem of learning is seen as that of predicting at one point (in the brain of the learner) either exactly or approximately an event that occurs at another point (in the environment). As in Shannon's view of communication, the formulation of these predictions is a reconstructive process based on a source code (prior experience) and the system is "designed" to predict not just events that have occurred, but also events that might occur (i.e., in learning, experience modifies the source code to enable more exact prediction of events in the future).

Do human beings act as ideal decoders?

Gallistel's results showing that animal learning can be modeled as decoding in information theoretic terms are encouraging if one's goal is to apply information theory to human communication and learning; however for all the encouragement they offer, they are no more than suggestive in this regard. It doesn't follow from that humans do act as ideal decoders, nor does it follow that human communication has the kind of mutually predictive structure that one might envisage from were one to view it from an information processing perspective. To return to Shannon (1956), the question of whether human beings can and do act like an ideal decoder in communication, "is an experimental and not a mathematical fact, and as such must be tested under a wide variety of experimental situations."

One initial test of this suggestion is provided by Ramscar, Yarlett, Dye, Denny & Thorpe (2010). Ramscar et al exploited the discreet characteristics of linguistic signs to explicitly contrast the predictions of the standard linguistic model with those of an ideal decoder in a series of computational and empirical studies of learning. To explain these results, it will be first helpful to go into a little more detail about one of the results provided by Shannon (1948).

Shannon defined *entropy* to represent the absolute limit on the best possible lossless compression of any communication, given certain defined constraints. If messages are encoded as a sequence of independent and identically distributed random variables, then Shannon's (1948) source coding theorem shows that, in the limit, the average length of the shortest possible representation that can encode the messages in a given alphabet is their entropy divided by the logarithm of the number of symbols in the target alphabet.

Entropy (*H*) is thus the average probability of a given random variable *N* with a range of value $n_1,\ldots,n_k$, and can be defined as:

$$H(N) = -\sum_{i=1}^{k} P(n_i)\log_2 P(n_i) \tag{3}$$

Entropy allows the expected value of the information contained in *N* to be quantified, and this quantity is usually given in units such as bits (where a bit is the amount of information stored by a digital device or other physical system that exists in one of two possible distinct states).
A fair coin has an entropy of one bit. However, if the coin is biased, then uncertainty about it will be lower (when asked to bet on the outcome of throws of a biased coin, one would preferentially bet on the most frequent result), and thus the Shannon entropy is lower. A succession of repeating characters will have an entropy rate approaching 0, since every successive character will become more predictable. Accordingly, the more uncertainty there is about *N*, the higher its entropy will be in bits, which means that the size of the shortest possible representation that can encode the messages in a given code will increase accordingly.

To get a better understanding of what this means in terms of human communication, if we assume that to learn and communicate about distinct entities or states of affairs in the world we must discriminate them, then entropy can usefully be understood as a quantification of discriminability, in that two different physical states are needed in order to distinguish between two possible entities in a code; and that differences in discriminability are balanced by biasing the coding of less likely (and therefore less easily discriminated) coded entities over more likely (and therefore more easily discriminated) entities, so that less likely entities are encoded more informatively than more likely entities (see also Ramscar et al, 2010).

The idea that every discrete piece of information *must* be encoded lies at the heart of information theory, and the underlying logic of this is particularly relevant to the idea that the relationship between signs and their meanings is bi-directional (so that the meaning of a sign can be inferred from it). As Ramscar et al (2010) point out, this idea (which is a basic assumption of the standard model of language) is at odds with the idea that symbols are *abstract* representations, because abstraction is not a bi-directional process. Abstraction involves reducing the information content of a representation, so that only information relevant to a particular purpose is retained (Hume, 1740; Rosch, 1978), which means that in Shannon's terms, abstraction is not lossless, and as such, cannot be a bi-directional process: one can abstract *from* a larger body of information *to* an abstract representation, but one cannot reverse the process simply because discarded information is—by definition—not coded for, and therefore unrecoverable.

Ramscar et al tested two hypotheses relating to the assumption of bi-directionality: first, that the sparse, discrete nature of signs (e.g., the word "dog") would not provide sufficient coding resources to discriminate between the elements of complex, high dimensional sorts of things signs actually tend to refer to in the world (such as dogs), and second, that if the coding relationship was reversed, so that objects in the world served as the basis for coding signs, then successful coding of meaning-sign relationships would be possible, and that the discrimination of the meaning of signs would be driven by abstraction (the discarding of unnecessary information).

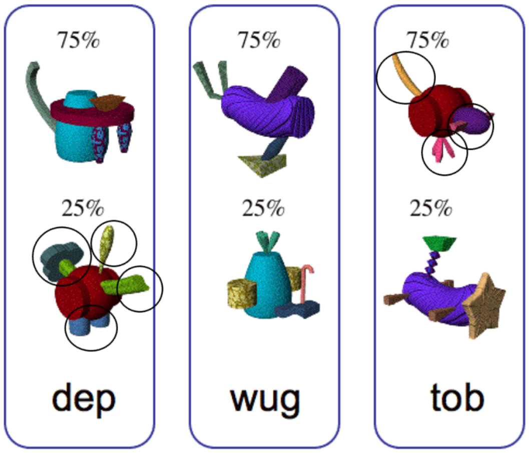

Figure 4. Examples of the category stimuli and structures trained in Ramscar et al (2010) Experiment 1 (James, Shima, Tarr, & Gauthier (2005). Note that body type does not discriminate between categories. The discriminating features that need to be learned in order to successfully distinguish the sub-categories are circled on the low-frequency "dep" and high-frequency "tob" exemplars. Ramscar et al found that participants were able do this if training was structured to allow for discrimination as a result of cue competition (i.e. if there were sufficient coding resources available). In this case, the shared high frequency "tob" and low frequency "dep" bodies caused participants to erroneously expect "dep" on "tob" trials, causing the error-rate of the bodies to rise relative to the other features that predicted "dep," and "tob." Cue competition then resulted in these other features being favored as predictive cues, and information about the body features being discarded. When training was structured to inhibit cue competition (i.e., when the number of coding bits was drastically reduced relative to the information to be

encoded), participants simply learned the degree to which each label predicted each feature. Participants who learned the category structures in this way failed to classify the low frequency exemplars.

To examine this suggestion formally, Ramscar et al (2010) trained participants on the category structures shown in Figure 4. To successfully learn to classify the objects in this task, participants had to learn to ignore (discard) salient yet uninformative features in forming a representation that would facilitate classification. In other words, success in the task required learning to discard information about the bodies of the objects because these were uninformative cues to category membership, which both required learning from prediction errors generated by the body cues, and recoding the residual predictive value that was subtracted from the body features to other cues.

Because learning is driven by prediction (expectation) and the structure of information (the number of coding bits available with respect to the amount of events that subsequently occur) can play a critical role in whether or not competitive learning occurs in a discriminative learning model. Learning the fribble features in Figure 4 as cues to discrete Labels – such as "wug" or "dax" (Feature-to-Label Learning) – allows for competitive learning amongst the co-varying features (which provide for a rich coding media), allowing value to shift from features that produce more error to those that produce less, discriminating the informative features from those that are uninformative (Ramscar et al, 2010). However, when this arrangement is temporally reversed, and the process becomes one of learning to predict a set of Features from a discrete Label (Label-to-Feature Learning), competition between cues cannot occur, since the label is the only cue available (value cannot transfer to other cues when there are none). Accordingly, when the label simply cues each feature at the level of its experienced frequency, what is learned is a model of the probability of every feature related to the label. Thus while FL-learning results in an informative model – cues are weighted according to their value in predicting the labels – LF-learning results in a simple probabilistic model of the contingencies of the learning environment (Ramscar et al, 2010).

By employing a speeded learning paradigm in which the presentation time of the objects to be learned about was limited to just 175 ms, Ramscar et al were able to control whether competition was available to participants in learning. Ramscar et al found that participants were only able to learn to categorize the low frequency exemplars in Figure 4 when FL-training allowed for the informative use of negative evidence in discriminative learning. Participants given LF-training failed to learn to correctly classify the low frequency exemplars.[2]

For the present purposes, there are two points worth making about this result: First, learning about the meaning of signs is clearly *not* a bi-directional process. When participants were forced to use their representations of the features of objects as a source for encoding labels, they learned the meanings of the signs very successfully; when the labels were made to act as a source for encoding features, they did not learn the meanings of the signs. Second, it seems that learning to code the relationship between signs and the things in the world they relate to does appear to involved discarding information—it is a lossy process—and thus if one's theory of communication requires that meaning be recoverable from signs (as the standard model does), then this is a problem.

If on the other hand, we assume that human communication is shaped to satisfy information theoretic constraints, and that language is a system in which meaning is something that a listener generates as part of the process of decoding a message (where decoding simply means accurately reconstructing a message with a high degree of confidence), then the fact that participants in Ramscar et al's experiment appeared to be highly adept at learning meanings as cues to signs is very promising, since it would suggest that this kind of learning is highly compatible with

[2] The effect of this manipulation is comparable to the difference between what are often termed generative and discriminative learning strategies, which have been much explored in machine learning. In machine learning, both generative and discriminative strategies are often used for estimating the likelihood that a particular item belongs to a given class in a classification task. Generative learning strategies solve this problem by building a probabilistic model of each category, and then using probabilistic inference to identify which class was most likely to have generated an item. By contrast, discriminative learning strategies seek to optimize the estimation of the probability of class labels, given the item to be classified. Studies have shown that generative and discriminative approaches differ in the speed and accuracy of their learning, the representations of the classes that they produce, and the kinds of classifications they result in (Efron, 1975; Rubinstein & Hastie 1997; Ng & Jordan, 2001; Bouchard & Triggs, 2004; Xue & Titterington, 2008; Ramscar et al, 2010; see Yuille 2005; 2006 for a discussion of conditions under which the predictions of the two converge).

Shannon's system model of communication, and the idea that humans can behave in a way that approximates an ideal decoder.

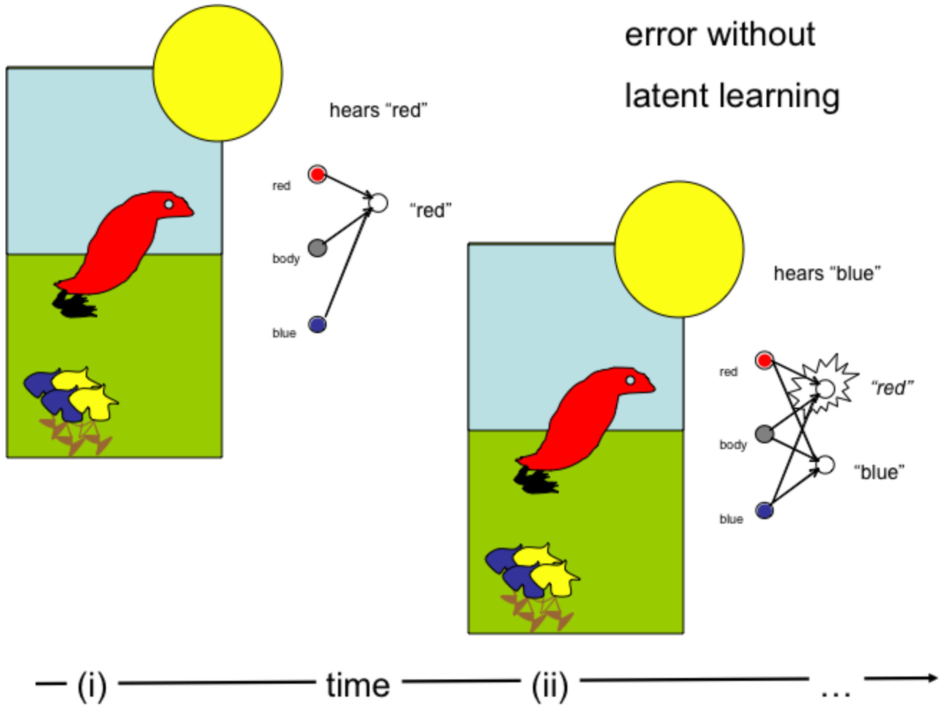

Figure 5. A depiction of the problem posed by color-word learning in a natural environment. In word-learning. Given the ubiquity of color, hearing a color word in isolation is uninformative, since red and blue are *both* likely to be present in the environment whenever "red" and "blue" are heard. To learn to map hues to color words, a child needs an information structure that favors the correct mappings: for example, provided the child knows the word for "wug," as "then look at the wug it's red" will provide an information structure that will allow a child to discriminate a meaningful mapping between the hue red and "red" (the information structure in this instance is equivalent to FL-learning in the adult experiment conducted by Ramscar et al – predicting labels from the world provide the child with a rich media for encoding the correct mappings to labels).

The drawback with this experiment is, of course, that it only vaguely approximates the kind of language-learning situations children are actually exposed to. Ramscar et al address this problem by showing how the same kind of analysis of information structure can be used to solve a problem first noted by Darwin (1877), children's color word learning:

> *"I carefully followed the mental development of my small children, and I was astonished to observe [that] soon after they had reached the age in which they knew the names of all the ordinary things that they appeared to be entirely incapable of giving the right names to the colors of a color etching. They could not name the colors, although I tried repeatedly to teach them the*

> *names of the colors. I remember quite clearly to have stated that they are color blind."* (Darwin, 1877).

Ramscar et al observed the ubiquity of colors of all kind may be part of the problem facing children. Because children will constantly be exposed to a variety of hues, hearing a color word in isolation is uninformative, since red and blue are both likely to be present in the environment whenever "red" and "blue" are heard. Ramscar et al found that if children knew the word for objects, training on sentences such as "look at the *object* it's red" provided an information structure that allowed child to discriminate a meaningful mapping between the hue red and "red" (the information structure here is equivalent to FL-learning in the adult experiment conducted by Ramscar et al – predicting labels from the world provide the child with a rich media for encoding the correct mappings to labels). However, when children were provided with sentences such as "look at the red *object* " failed to learn to map hues to color label, because in this instance, hearing the color word prior to the object provided only an insufficient medium for encoding the information required to make the appropriate mapping.

To further examine whether children's learning to map meanings to signs approximates an ideal decoder, Ramscar, Dye, Muenke-Popick & O'Donnell-McCarthy (2011) gave children training in which they were shown a set of objects (say bears), and asked, "What can you see? Bears. There are four." As a result of this, Ramscar et al expected children would learn about the relationship between the pictorial representation (the bears) and the number word (four). Specifically, on a trail with theis information structure, the child will learn:

1. All of the features present (e.g., color, shape, set-size, etc) will be reinforced as cues to "four."

2. Given that the child is playing a 'number' game, in which number words have been generated from similar items, she will be expecting to hear other numbers as well (such as "three" and "five"), based on prior experience. This will generate an error signal for all of the features that prompted those erroneous expectations, which will cause the features present to be unlearned as cues to other number words. This means

> that not only will set-size four be unlearned as a cue to “three” and “five,” but so will all the consistently unreliable features, such as color, shape, and so on.

Thus, while 3, 5 and 7 were not trained in the experiment, if children were learning to make informative predictions, then 2, 4 and 6 trials will provide information about 3, 5 and 7, even when there is no explicit training on 3, 5 and 7. For example, while the individual cue value of set-size five cannot be affected by learning on a 4-trial, since it is not present, other potential predictors of set-size five will be present on that trial (e.g., color, shape, and so on), and their informativity as cues to “five” will be affected by this. The non-occurrence of “five” will cause the informativity of cues that erroneously predicted “five” to decrease, which will result in a relative increase in the informativity of set-size five as a predictor of “five.” As a result, the child will be able to better discriminate the cues to “five” (i.e., she will have learned about “five”), even though – or, in fact, because – “five” was not heard in this context.

Perhaps counterintuitively, this also means that for a child who has some experience of “four,” but has yet to fully discriminate the mapping between set-size four and “four,” trials in which “four” is presented will not be particularly informative about “four.” This is because the informativity of set-size four will not change relative to any other information present on 4-trials, since they will all be similarly learned as informative predictors of “four.” On the other hand, a 4-trial will decrease the informativity of any cues that prompt erroneous expectation of other number words. Thus, for example, an object’s shape might lose its informativity about “five,” making set-size five more informative about “five” (and so on), and over time this process should help a child improve her discrimination of a system of numbers.

Consistent with this analysis, Ramscar et al (2011) found that FL-trained children not only improved in their ability to identify sets of 2, 4 and 6, but also that they improved on discriminating 3, 5 and 7 when none of 2, 4 and 6 were present. Children given LF-training (“What can you see? There are *two* balls”) showed no improvement on any of the tested set sizes. (Similarly, Ramscar & Dye, 2011, showed that training older children solely on regular plurals could lead to a *reduction* in plural over-regularization, for much the same reasons.)

Finally, Ramscar, Suh & Dye (2011) applied the same approach in an examination of the reasons for why people tend to lose “perfect” or “absolute pitch” – the ability to name notes based on absolute frequency information. It has long been known that 1) perfect pitch is typically quite rare among the general population and 2) that possessors perform less well than non-possessors on tests of relative pitch. Ramscar, Suh & Dye (2011) used an FL-LF design to train participants to either acquire more accurate representations (LF) – which, since it can only produce a generative model of the tone space, ought to be more informative in absolute pitch tasks – or better between-category representations (FL) – which, since it increases the discrimination of categories, ought to be better for relative pitch learning, at a cost to accuracy. Consistent with this, Ramscar, Suh & Dye (2011) found that LF trained participants were better at perceptually discriminating trained notes from near-identical lures than FL-trained participants (indicating that they had formed a less abstracted representation of what they had heard), they performed proportionally less well at discriminating octave transpositions of the trained notes from lures (indicating that they had abstracted the pitch categories less well). A similar pattern of results was obtained in tasks matching tones to labels. Importantly, FL trained participants showed a loss of sensitivity to within-category discriminations as a result of learning pitch category representations. In other words, their representations of pitch had become more symbolic.

This latter finding thus further supports the idea that abstraction is not a lossless process. Further, it also suggests that signs themselves are probabilistic entities: signs can be seen as being more or less “symbolic” (and abstract) depending on the degree to which the information in the representations learners have for them has been shaped by abstraction.

Encoding, decoding and information loss

It is worth reflecting briefly on what the findings reviewed above mean for the standard model. The standard model of language (which depending on one’s take on the history of ideas, we inherited from either de Saussere, the Vienna Circle, or the Greeks) assumes that meaning is conveyed by language in much the same way that passengers are conveyed on a plane. The material to be conveyed (be it people or meanings) are squeezed in at the point of departure, and then unpacked at the point of arrival. The idea, formed before we had the formal concepts, is that linguistic coding is in some sense *lossless* (in communications engineering, lossless data compression is a class of data compression algorithms that allows the exact original data to be reconstructed from the compressed data). However, a key requirement for lossless data compression is that sufficient coding resources are deployed to allow all of the physical state differentiations in the original data set to be preserved and reconstructed (the original data is compressed by eliminating redundant or improbable information in the original data; Salomon & Motta, 2009). Yet the reason that very marked differences were seen in FL versus LF learning across all of the experiments reviewed above was because FL learning is lossy (Figure 6). People failed to discriminate labels after LF training *because* they build up codes that were in some sense accurate in their representation of the world, but uninformative about the labels they were supposed to be learning. On the other hand, people performed very well when it came to discriminating and using labels after FL training, but the reason that they were able to do so in that they learned abstract codes that were optimized for predicting the signs, and because this optimization process involved discarding any data that was uninformative to that task.

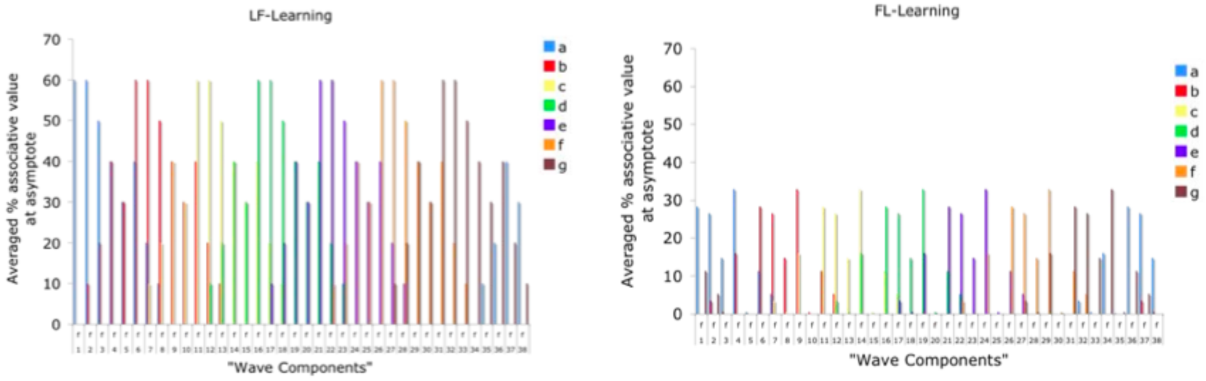

Figure 6. Simulations of how identical sets of labeled categories in overlapping artificial "waveforms" structured are learned when trained either LF (labels predicting waves) or FL (waves predicting labels; Ramscar, Suh & Dye, 2011). The categories are labeled a-g, and the probability of a wave component occurring in each category is represented in the right panel (the LF-learned model; because LF-learning results in a generative model, it accurately depicts the training data). As can be seen, the representation of each component in the FL-learned model differs markedly from its empirical rate of occurrence, with some components being completely discarded in order to better discriminate the code that best predicts the labels. Critically, the FL-learned model does not contain sufficient information to allow the LF-model

to be reconstructed from it.

This in turn suggests two things: First, if we agree that the fundamental problem of *communication* is that of reproducing at one point either exactly or approximately a message selected at another point, then human learning appears to be ideally suited to learning source codes that will allow for the construction of messages about the world. However, while it is certainly true that these messages will refer to or be correlated according to some system with certain physical or conceptual entities, it would appear that the design of human learning considers these semantic aspects of communication largely irrelevant to the engineering problem (most of this semantic information appears to be thrown away in coding). That is, it would appear that, supporting Shannon's (1956) suggestion, when tested under a wide variety of experimental situations, human beings do, in many situations, act like ideal decoders. Second, it would appear that language does not convey meanings in the way that is commonly supposed. One of the basic ideas of the standard model of language—that meanings are coded for in speech or text, and can be extracted from them—appears to be highly inconsistent with the way human beings actually learn and use linguistic codes (this problem has been noted before, see e.g. Wittgenstein, 1953; Quine, 1960; the only difference in this instance is that information theory allows it to be formalized and examined). Moreover, not only does it appear that appeals to innate abilities won't seem to help here, but further, the empirical facts militate against the idea that genetically specified "concepts" can still (somehow) allow meanings to be encoded in and decoded from linguistic signals. (It is true that one might defend the standard model by appealing to pragmatic inferences at this point, but this in turn raises the question of why one would assume a model based on the encoding decoding of meaning in signals if meanings are not actually decoded from signals, but rather generated by pragmatic inference.)

Some properties of the code

Like science, I will argue that linguistic communication is a conventionalized, probabilistic process aimed at understanding, and that it is driven by evidence. Speakers are engaged in the business of predicting one another, and the more evidence they can amass (based on linguistic, social, conventional, etc, factors) the better they will understand one another. While this view is in many ways at odds with the standard model of language, there are a number of reasons for believing that there may be the benefits to be had from exploring alternative conceptions of language and communication.

Prior to the middle of the 20th century, numerous alternative approaches to the study of language were being explored (e.g., Wittgenstein, Skinner, Quine, Harris), and the "revolution" that led to the dominant hold that varieties of the transformational, generative account of language have on theory was not founded on any empirical or theoretical *discoveries*, but rather on claims about the inadequacies of other approaches (Chomsky, 1957; 1959; Miller & Chomsky, 1963). The reason so many people have accepted what should be the unacceptable conclusion that language and meaning are innate is because of a widely held belief that language could *only* be explained by some kind of generative account, and that, in principle, there can be no alternative conception of language.

Some of the inadequacies of these arguments have already been laid bare. For example, Chomsky (1957) argues *against* the kind of predictive account embodied in an information theoretic approach to language by noting (correctly) that *Colorless green ideas sleep furiously* and *Furiously sleep ideas green colorless* have the same frequency in English (before 1957 this was zero). Based on this, Chomsky makes the empirically false claim that, "Therefore, [we are] obliged to conclude that frequency reveals nothing about grammaticality." As Pereira (2000) shows, this erroneous conclusion relies on the assumption that the probability of a type of event must be regarded as zero if it has not occurred so far (a result of using maximum likelihood estimation, MLE), and that if a better estimation method is applied ('smoothing,' Good, 1953; McKay, Degan, Lee & Perreira etc.) then based on normal English text, the first sentence is actually 200,000 times more probable than the second.

Similarly, Miller & Chomsky (1963) offered an analysis that is taken by many as providing a demonstration that providing a probabilistic account of language is impossible. Miller & Chomsky observed that for an n-gram model to represent the intricacies of even a moderate proportion of English sentences, an apparently unlearnable number of statistical parameters would need to be estimated:

> "Just how large must n and V be in order to give a satisfactory model? Consider a perfectly ordinary sentence: The people who called and wanted to rent your house when you go away next year are from California. In this sentence there is a grammatical dependency extending from the second word (the plural subject people) to the seventeenth word (the plural verb are). In order to reflect this particular dependency, therefore, n must be at least 15 words. We have not attempted to explore how far V can be pushed and still appear to stay within the bounds of common usage, but the limit is surely greater than 15 words; and the vocabulary must have at least 1,000 words. Taking these conservative values of n and V, therefore, we have $V^n = 10^{45}$parameters to cope with, far more than we could estimate even with the fastest digital computers." (p. 430; We have altered Miller & Chomsky's notation from K and d to V and n for clarity.)

Thus, as n increases, the number of potential parameters to be estimated grows as V^n, where V is the number of tokens in the language. We can contextualize this (conservative) estimate of the number of parameters that any language model might need to estimate in various ways. For example, it has been estimated that a child with professional parents will only be exposed to around 11 million spoken words a year (or 11×10^6; Hart & Risley, 1995), and that the average lifetime consists only of around 2.2×10^9 seconds. By either measure, the figure of 10^{45} put forward by Miller and Chomsky is almost unimaginably larger.

Once again, Miller and Chomsky's assumption relies on MLE. Methods such as Good-Turing estimation allow for the estimation of the probability of hitherto unforeseen events (Pereira, 2000), and unless we are to accept that not only syntax and semantics, but also our entire life histories are innate, the ability that people have to both learn from experience, and apply what they have learned in new circumstances, suggests that they too are capable of estimating the probability of novel events. Further, as Ramscar et al (2010) note, the wealth of experimental evidence showing that people are *highly* sensitive to the statistical properties of their languages would appear to offer an empirical refutation of Miller and Chomsky's claim (unless, that is, we assume that people's knowledge of the statistical properties of the vocabulary items in their languages is innate too).

A second class of negative claim that has historically been used to undermine alternative approaches to human communication are arguments to the effect that a child learning language is somehow starved of useful information, and that language learning is impossible (poverty of the stimulus arguments). In a series of helpful and astute reviews, Pullum & Scholz (Pullum & Scholz, 2002; Scholz & Pullum, 2002; Scholz & Pullum, 2006) show how the literature on child language learning is awash with claims that generative grammars cannot be learned from the kind of language that a child is exposed to:

> [poverty of stimulus arguments take] the following form: (i) a fact about some natural language is exhibited that allegedly could not be learned from experience without access to a certain kind of (positive) data; (ii) it is claimed that data of the type in question are not found in normal linguistic experience; hence (iii) it is concluded that people cannot be learning the language from mere exposure to language use (Pullum & Scholz, 2002, p. 9).

In showing that many such arguments fail to establish their claims, Pullum & Scholz note that another set of conclusions can be drawn from such arguments: namely (iv), that the linguistic “fact” that is supposed to be unlearnable is actually nothing of the sort, or (v), that learners don’t solve the learning problem in the way that a given argument supposes. Thus, given a poverty of the stimulus argument, it is possible *either* to conclude that language is innate, *or else* that the theoretical assumptions that lead to a nativist conclusion are false.

The development of technology that allows large electronic samples of linguistic data to be stored and analyzed provides ever more support for these latter conclusions. For example, Baayen et al (in press) show that if the task of language learning is analyzed in terms of discrimination (i.e. that the goal of learning is to approximate a decoder, in Shannon’s terms), a modified Rescorla-Wagner implementation (Danks, 2003) trained on a relatively small corpus (~11 million two- and three-word phrases) can provides good empirical fits to human data on a wide range of effects documented for lexical processing, including frequency effects, morphological family size effects, and relative entropy effects. For monomorphemic words, the model provides excellent predictions with no free parameters, and for morphologically complex words, Baayen et al had only to add a few free parameters to enable the model to fit a broader range of data better and more parsimoniously than other models in the literature that were designed specifically for the task (e.g., Norris, 2006). Further, the model also captures frequency effects for complex words, and phrasal frequency effects, even though it has no explicit representations of complex words or phrases (Baayen and Hendrix, 2011): it simply learns whatever discriminative (informative) code is in the training set, and then reuses this when tested.

In other words, Baayen et al discovered that training a reading model on a small, representative sample of language enabled that model to respond to new data in a way that closely approximated a human reader. This suggests that far from being impoverished, the information in that a small, representative sample of language provided a good model of at least some human linguistic knowledge, and it hardly seems pushing the bounds of credulity to suggest that if a model based on empirically derived principles of learning can gather that kind of information from the linguistic environment, actual human language users will be perfectly capable of learning that information as well.

Perhaps even more suggestively, Ramscar, Futrell & Dye (2011) have shown that the German grammatical gender system—derided as meaningless for generations by linguists, philosophers and psychologists—is an exquisitely structured system than balances information and learning constraints to manage the entropy of nouns in context. have accepted the idea that grammatical gender may have evolved separately in many different languages, even though it serves no apparent purpose. We question this assumption by considering whether German gender might play an informative role in language after all. By analyzing the information structure of German, we reveal how gender serves to make nouns more predictable in context. Ramscar, Futrell & Dye show not only that the gender system has structure, but also that this structure mirrors that of other subsystems of language (such as verb inflection), in that it is more specifically informative about high frequency items than lower frequency items (from the point of view of learnability, it makes obvious sense that information about high frequency items can be encoded in more efficient, conventionalized manner than that for than lower frequency items, simply because the probability that all the speakers within a given community will have the opportunity to master a convention will rise with its frequency).

Ramscar, Futrell & Dye show that in comparison to English – a Germanic language that has largely shed its gender system – grammatical gender allows German speakers to use a wider variety of nouns after articles. They then show how English has systematically compensated for its diminished gender system by extending the use of pre-nominal adjectives, employing them with greater frequency as the frequency of the nouns they precede decreases. Not only do English pre-nominal adjectives help to make nouns more predictable in context, but Ramscar, Futrell & Dye show that the distribution of pre-nominal adjectives is organized in the linguistic environment so that it *optimizes* this function, ensuring that pre-nominal adjectives provide more support for low frequency nouns than high frequency nouns, thereby helping to make all nouns equally predictable in context.

These findings are consistent with Zipf's famous 'Principle of Least Effort,' which holds that human behavior is shaped by a bias to minimize people's "average rate of work-expenditure over time" (Zipf, 1935; 1949; this might also be thought of in terms of channel capacity, *pace* Shannon, 1948). Indeed, a growing body of evidence supports the idea that in language use, people manage the rate at which information is encoded in linguistic signals, avoiding excessive peaks and troughs in entropy across messages (Aylett & Turk, 2004; Levy, 2008; Jaeger, 2010).

For example, Genzel and Charniak's (2002, 2003) propose and provide support for the *entropy rate principle*, which holds that users produce language in a way that maintains a relatively constant average entropy rate. The motivation for this principle comes from information theory, which suggests that the most efficient way of transmitting information through a noisy channel is at a constant rate. The principle is empirically grounded, with a raft of findings suggesting that it holds true in speech and in text. For instance, in text the entropy rate principle predicts that if the sentences of a given text are equally informative when encountered in context, *ignoring* context should cause entropy to steadily increase as a text progresses (because the meaning in earlier parts of a text will serve to generate an informative context that will reduce the entropy of later parts). This finding has been confirmed in numerous corpus studies, and the same entropy-increasing pattern has been found across different genres, and even within paragraphs (Genzel and Charniak, 2002; 2003; see also Keller, 2004; Levy, 2008; Jaeger, 2010).

Similarly, in speech, word length – as measured in phonemes or syllables – correlates better with a word's average predictability in context than with its frequency (Manin, 2006; Piantadosi, Tily,

& Gibson, 2011). This suggests that the more probable a word is in context, the less information it carries, and the more redundant it is, resulting in a shorter phonological form. In the same vein, studies have shown that even for the same word, instances that are more predictable will be produced with shorter duration and with less phonological and phonetic detail (Aylett 1999; Bell et al., 2003; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Pluymaekers, Ernestus, & Baayen, 2005; van Son et al., 1998; van Son & van Santen, 2005).

Taken together, these findings suggest not only that the linguistic environment is far from impoverished, but further, that viewed appropriately—as a source of information for predicting linguistic signals, rather than abstracting an ill-defined "grammar"—the linguistic environment is a rich repository of information for a learner trying to master the source code for her language.

Human communication: the system solution

If we allow ourselves to entertain the idea that there might be a better alternative to the standard model of language is wrong, this in turn poses a question: what is it? I suggest the solution is the same for human communication as it is for other communication systems, namely the system solution of Shannon and Hartley. Whereas nativist approaches within the generative linguistic tradition (which also includes most of psychology over the past 50 years) have focused on the physical system (imbuing children with more and more innate abilities to facilitate communication), the evidence from both communication theory and psychology points in another direction: the solution must lie, in large part, in the system.

Just as information theory sees, "the fundamental problem of communication [as] that of reproducing at one point, either exactly or approximately, a message selected at another point." (Shannon, 1948), the problem of explaining the communication of semantic information can be seen as explaining how a listener reproduces a speaker's intended meaning (Ramscar et al, 2010). From this perspective, linguistic communication can be seen as a probabilistic process in which a speaker helps a listener to predict, either exactly or approximately, the speaker's intentions (Ramscar, 2010; Ramscar et al, 2010; Ramscar & Dye, 2009; Ramscar & Yarlett, 2007; Baayen, 2011; Baayen & Hendrix, 2011; Baayen et al, in press).

What does this mean? First and foremost, it means that understanding language in terms of information means that we should give up the idea that language users decode, or extract, meaning from linguistic signals. The fundamental problem of the human communication system faces is that same as any other communication system: reproducing at one point either exactly or approximately a message selected at another point. That these messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities *is* irrelevant to this engineering problem. The significant point to grasp is that the problem is that of selecting the actual message from a set of possible messages. The point—and it is a very large one—is that a human learner will only be able to select the appropriate message if they have learned the source code, and are able to use context to help in this process paragraphs (Genzel and Charniak, 2002; 2003). That is, users don't decode meaning from linguistic signals, but rather they build up meanings in the process of decoding them: meaning is something a listener or reader constructs out of the source code she has learned, and the context provided by the next part of a message, to predict upcoming parts of a message.

To return to an example of abstraction discussed above, the idea that a meaning can be conveyed by a word makes no more sense than the idea that someone might be able to "get" detailed information about the results and methods sections of a paper they had never seen, simply by reading its abstract. Given an abstract, one can only make guesses about the results and methods sections, making a kind of prediction about the kind of information they might contain. If the reader is an expert, the likelihood that these predictions will be more accurate, or even substantially correct, will increase. However, *given no more than an abstract*, *the reader can do no more than make predictions*, because the process of abstraction involves discarding information that cannot be later recovered from an abstract representation. Symbols are abstractions, and it follows similarly that the meanings of signals can only be inferred; language is a probabilistic practice in which a speaker helps a listener to predict, either exactly or approximately, the speaker's intentions (Wittgesntein, 1953; Shannon, 1948).

While this view is at odds with the standard view of language, it is highly compatible with an enormous array of empirical findings relating to language processing and the nature of language.

Numerous results in a variety of research paradigms have revealed that when people are listening to or reading a sentence they build up a rich set of linguistic *expectations*, predicting upcoming words based on the structure and semantics of the prior discourse (see e.g., Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995; De Long, Urbach & Kutas, 2005; Kutas & Federmeier, 2007; Pickering & Garrod, 2007; Garrod & Pickering, 2009; Tanenhaus & Brown-Schmid, 2008; Norris & McQueen, 2008; Altmann & Mirković, 2009). While this may seem obvious in idiomatic phrases, such as "*cross my heart and hope to* ___" or "*hit the nail on the* ___" a considerable body of evidence is consistent with the idea that this is true of language more generally, and that it is reflective of the fact that comprehenders behave like ideal decoders.

For example, De Long, Urbach & Kutas (2005) measured event related potentials (ERPs) while participants read sentences like "*the day was breezy so the boy went outside to fly a kite"* and *"the day was breezy so the boy went outside to fly an airplane."* Not only did they find that the less predictable *airplane* produce a larger n400 than *kite* (n400 is an ERP component typically elicited by an unexpected linguistic stimulus), but the same pattern held for the articles *a* and *an* as well. This suggests that participants were using context to predict not just *kite*, but also the article that preceded it, causing them to find "*an"* surprising as a result.

While studies to date have often focused on anticipation of a specific word, object or event based on prior context (see e.g., Altmann & Steedman, 1988; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; De Long, Urbach & Kutas, 2005; Otten & Van Berkum, 2008), it is clear that in natural speech, listeners are anticipating (probabilistically) a range of different possible words that might follow in a given speech stream (Norris & McQueen, 2008). In much of the literature these predictions are seen as *assisting* the meaningful production and comprehension of language, but from the information theoretic perspective described here these predictions can be seen as actually *comprising* the meaningful use of language, because the goal of the comprehender is simply to predict the signal (Shannon, 1948; Ramscar et al, 2010; Ramscar, Matlock & Dye, 2010).

There is a widespread consensus that "intention reading"— social prediction—is an important component of word learning (see e.g., Bloom, 2000; Tomasello, 2003; 2007); our proposal is that "intention reading" can be extended to language processing more generally. We argue that comprehension arises both out of what the listener *knows*—what she will predict—and what the listener *is learning*—what she will come to predict. In order to predict a speaker, even partially, a listener must activate similar cues to those used by the speaker in generating an utterance. As the speaker uses FL-learned cues to generate speech, the listener will activate similar FL-learned cues, allowing her to anticipate upcoming speech. When the listener does not fully predict the speaker's words—as will often be the case—further FL-learning will take place. Comprehension, like learning, is thus a process of having and modifying expectations.

The most likely way a listener can predict a speaker is by sharing a cultural "source code" (as well as learning to fit this locally, c.f. Davidson, 1986). Linguistic distributions are not structureless, but rather, then might be better seen as a repository of recipes for the expression of ideas, shaped by the conflicting needs of learners and communicators. Communities of learners and communicators, their interactions and their different needs are the forces that shape the social evolution of languages; and the evolutionary solutions that these communities evolve become encoded in the specific patterns of words they speakers use and transmit to learners. These patterns are a kind of or a kind of social art form, embodying the communicative recipes that each community has unconsciously evolved.

To give an example of this, English speakers demonstrate a remarkable tendency to describe puppies as *cute, little.* Moreover, they are more likely to use redundant adjectives in conjunction with more specific, lower frequency nouns (Ramscar, Futrell & Dye, 2011). From the perspective of the standard model, this behavior is somewhat puzzling, given puppies are cute and little, and given people know what *cute, little* and *puppy* mean, it isn't at all clear *why* people bother to say, "cute little puppy," at all, let alone why it is that they do so with quite amazing consistency.

If language is seen as an informative, discriminatory process, and if the distribution to which people have been exposed is seen as a source code, on the other hand, then the reason why people say things like, "cute little puppy," becomes apparent: not everything is cute, little, or a puppy, and thus "cute little puppy," is a communicatively efficient way of communicating about puppies, since these adjectives iteratively reduce uncertainty in messages about puppies. In the same vein, it makes more sense for someone rummaging in a refrigerator to ask if a guest would like “a cold beer” rather than a “crisp beer” (even though it should be obvious that if the beer is in the fridge, it is going to be cold, and even where the beer in question is a particularly crisp little brew) because “cold beer” is highly conventionalized in English (in the 400+ million word COCA corpus, Davies, 2009, cold and beer have a Mutual Information of 6.59) whereas “crisp beer” is not (there are no instances in the 410 million word sample in COCA). Thus, by saying “cold beer,” the host will increase the likelihood that her guest is able to understand her, and the chances, too, that she will not have to repeat herself. “Cold,” in this context, is semantically redundant but highly informative, and we would suggest that it may not take too much of a leap to see how this kind of usage, when highly conventionalized, could grammaticalize over time, leading to the perplexing system so hilariously (and unfairly) lampooned by Twain.

These examples show how in a communication system, meaning will be shared to the degree that predictions are shared. The communication of meaning can thus be seen as a process of aligning a speaker and a listener’s predictions (see also Davidson, 1986; Pickering & Garrod, 2007; Garrod & Pickering, 2009). Successful communication is a process of reducing uncertainty in a listener’s predictions as an utterance or dialog unfolds. We should note, that “predicting,” does not mean “predicting with certainty.” From this perspective, understanding is inherently probabilistic: when the degree of prediction is too low, little communication is occurring, and the listener is likely confused; when prediction is too high, communication is occurring, however, the listener is likely deeply bored.

If human communication involves predicting signals—and I have outlined many reasons for assuming that it does—then meaning is something that a speaker elicits in a listener simply by engaging the listener in a game of prediction. In this game, signals aren’t used to *convey* meaning, but rather are used to reduce a listener’s uncertainty about a speakers’ intended message (Shannon, 1948). In order for a listener to predict a speaker, the listener has to activate the same semantic cues to the signal as the speaker, such that the listener comes to understand an utterance by *thinking* about that utterance in a way that converges on that of the speaker. This proposal has much in common with the idea that language is a form of joint action (see, e.g., Wittgenstein, 1953; Quine, 1960; Pickering & Garrod, 2007; Garrod & Pickering, 2009; Tanenhaus & Brown-Schmid, 2008; Altmann & Mirković, 2009; Gennari & MacDonald, 2009); it differs only in that it explicitly frames communication as a systematic, information-based process. A proper understanding language in terms of information may involve a reassessment of what human communication involves, and it will certainly require a large revision of our theories of language and its role in culture (Wittgenstein, 1953; Quine, 1960; Tomasello, 1999, 2003; see also Fodor, 2000); it will change, and may well help the way we talk to one another, and the ways in which we collaborate to make meaning.

References

Altmann, G. T. M., & Mirkovic´ , J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33, 1–27.

Arnon, I., & Ramscar, M. (2009). Granularity and the acquisition of grammatical gender: How order-of- acquisition affects what gets learned. *Proceedings of the 31st Annual Conference of the Cognitive Science Society,* Amsterdam

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1), 31–56

Baayen, R. H., & Moscoso del Prado Martin, F. (2005). Semantic density and past-tense formation in three Germanic languages. *Language*, 81, 666–698

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P. and Marelli, M. (2011), An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*.

Baayen, R. H. and Hendrix, P. (2011) Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. *'Empirically examining parsimony and redundancy in usage-based models*', Linguistic Society of America

Baayen, R. H. (2011) Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*

Bell, A., D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory & D. Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113(2). 1001-1024

Bloomfield, L. (1933). *Language*. Revised from 1914 edition. New York: Holt

Chomsky, N. (1957). *Syntactic structures*. The Hague, The Netherlands: Mouton.

Chomsky, N. (1959). Review of verbal behavior, by B.F. Skinner. *Language*, 35, 26–57.

Chomsky, N. (1997) *New Horizons in the Study of Language and Mind* Cambridge, England: Cambridge University Press.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech*. Cognitive Psychology*, 37, 201-242

Claudi, U (1985). *Zur Entstehung von Genussystemen: Überlegungen zu einigen theoretischen Aspekten, verbunden mit einer Fallstudie des Zande: Mit einer Bibliographie und einer Karte.* Hamburg: Buske)

Corbett, G. G. (1991). Gender. Melbourne: Cambridge University Press.

Craig, Colette. 1986. "Jacaltec noun classifiers. A study in grammaticalization", *Lingua* 70, 4: 241-284.Craig 1986

Curzan, A. (2003). *Gender shifts in the history of English.* Cambridge: Cambridge University Press.

Dawson, H. C. (2003): Defining the Outcome of Language Contact: Old English and Old Norse. *OSUWPL* 57: 40-57

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language*, 42, 465-480Davies, 2009

Eisenberg, P. (2006). *Grundriss der deutschen Grammatik*. Metzler Verlag.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8, 1117–1121

Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. New York: Oxford University Press.

Fodor, J. (2000). *The mind doesn't work that way: The scope and limits of computational psychology.* Cambridge, MA: MIT Press. .

Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment and dialogue. Topics in Cognitive Science, 1, 292–304

Genzel, D. and E. Charniak (2002). Entropy rate constancy in text. In Proceedings of ACL-02.In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL '02). Ann Arbor, Michigan: Association for Computational Linguistics*

Genzel, D., & Charniak, E. (2003). Variation of entropy and parse tree of sentences as a function of the sentence number. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 65-72, Sapporo

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40 (3 and 4), 237-264

Hart, B., & Risley, T. (1995). Mea*ningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes

Heath, Jeffrey (1975). Some functional relationships in grammar. *Language* 51: 89-104

Hopper, P., & Traugott, E. 1993. *Grammaticalization*. Cambridge: Cambridge University Press.Hudson-Cam & Newport, 2009)

Jaeger, T.F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1), 23-62

Jurafsky, D, &. Martin, J.H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall.Kamp & Partee, 1995;

Karins, K., MacIntyre, R., Brandmair, M., Lauscher, S. & McLemore, C. (1997). CALLHOME German Lexicon. Linguistic Data Consortium, Philadelphia.

Kamp, H and Partee, B (1995) "Prototype theory and compositionality", *Cognition* 57, 129-191

Keller, F. (2004). The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 317–324, Barcelona

Kilarski, M. 2007. On grammatical gender as an arbitrary and redundant category. In Douglas Kil- bee, editor, *History of Linguistics 2005: Selected papers from the 10th International Conference on the History of Language Sciences (ICHOLS X),* pages 24–36. John Benjamins, Amsterdam

Klein, D. & Manning, C. (2003). Accurate unlexicalized parsing. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, *Ann Arbor, Michigan: Association for Computational Linguistics*

Koval', A. I. (1979). O značenii morfologičeskogo pokazatelja klassa v fula. In N. V. Oxotina (ed.) Morfonologija i morfologija klassov slov v jazykax Afriki, 5-100. Moscow: Nauka

Köpcke, Klaus-Michael (1982). *Untersuchungen zum Genussystem der deutschen Gegenwartssprache.* Tübingen: Niemeyer (Linguistische Arbeiten 122)

Kuhn, T.S. (1957) *The Copernican Revolution: planetary astronomy in the development of Western thought*. Cambridge: Harvard University Press.

Kuhn, T.S. (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kullback, S. & Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics* 22 (1): 79–86

Kutas, M., & Federmeier, K. D. (2007). Event-related brain potential (ERP) studies of sentence processing. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 385–406). Oxford, England: Oxford University Press

Lakoff, George (1986). Classifiers as a reflection of mind. In *Typological Studies in Language* 7: Noun Classes and Categorization (ed. Colette Craig), pp. 13-51

Levy, R (2008). Expectation-based syntactic comprehension. *Cognition* 106(3):1126-1177

Lewis, D.K. (1986) *On the Plurality of Worlds*, London: Blackwell.

Lezius, W, Reinhard, R. & Wettler, M. (1998). A freely available morphological analyzer, disambiguator, and context sensitive lemmatizer for German. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Vol. I–II . Montreal: Université de Montréal

Liberman, M (2011) Quoted in B Zimmer, On Language: The Future Tense*, The New York Times Sunday Magazine*, February 27, 2011, p MM16

Lupyan G & Dale R (2010) Language Structure Is Partly Determined by Social Structure. *PLoS ONE* 5(1): e8559

Manin, D. (2006). Experiments on predictability of word in context and information rate in natural language. *Journal of Information Processes*, 6, 229-236

Maratsos, M. P. (1979). How to get from words to sentences. In D. Aaronson & R. Rieber (eds.), *Perspectives in psycholinguistics*. Hillsdale, N.J.: Erlbaum

McDonald, S. A. & R. C. Shillcock (2003). Eye movements reveal the on-line computation of lexical probabilities. Psychological Science, 14, 648-652

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. II, pp. 419–491). New York: John Wiley

Newport, E.L. (1990). Maturational constraints on language learning. *Cognitive Science* 14: 11-28

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2), 357–395.

Partee, Barbara H. 2009. The dynamics of adjective meaning. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue 2009*" (Bekasovo, May 27-31, 2009), ed. A. E. Kibrik et al, 593-597. Moscow: Russian State Humanities University

Payne, J., Huddleston, R., & Pullum, G.K. (2010) The distribution and category status of adjectives and adverbs, *Word Structure* 3.1: 31–81

Pereira, F. (2000). Formal grammar and information theory: Together again*? Philosophical Transactions of the Royal Society* 358, 1239–1253.

Piantadosi, S.T., Tily, H., & Gibson, E. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*. 108(9):3526-9

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11, 105–110

Plag, I., & Baayen, R. H. (2009). Suffix ordering and morphological processing. *Language*, 85, 106–149

Pluymaekers, M., Ernestus, M. and Baayen, R. H. (2005) Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62, 146-159

Pullum, G. K. & Scholz, B. C. (2002) Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19, 9-50.

Ramscar, M. (2010) Computing Machinery and Understanding. *Cognitive Science* 34(7): 966–971

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010) Feature-label-order effects and their implications for symbolic learning. *Cognitive Science*, 34(7): 909–957

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Science*, 11(7), 274–279

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, 31, 927–960

Ramscar, M., & Dye, M. (2009). Expectation and negative evidence in language learning: The curious absence of mouses in adult speech. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Amsterdam

Rescorla, R.A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43, 151-160

Rescorla, R.A., & Allan R. Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H., and Prokasy, W. F. (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Croft

Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. B. Lloyd (Eds*.), Cognition and categorization* (pp. 27–48). Hillside, NJ: Lawrence Erlbaum Publishers

Salomon, D & Motta, G (2009) *Handbook of Data Compression*, 5th edition, Springer

Scholz, B. C. & Pullum, G. K. (2002) Searching for arguments to support linguistic nativism. *The Linguistic Review* 19, 185-223.

Scholz, B. C. and Pullum, G. K. (2006) Irrational nativist exuberance. In Robert Stainton (ed.), *Contemporary Debates in Cognitive Science*, 59-80. Oxford: Basil Blackwell

Shannon, Claude E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal,* 27: 379-423 and 623-65

Sigurd B,; Eeg-Olofsson M,; van de Weijer J. (2004) Word length, sentence length and frequency–Zipf revisited. *Studia Linguistica* 58:37–52

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American sign language from inconsistent input. *Cognitive Psychology*, 49, 370–407

Skut, W., B. Krenn, T. Brants., & H. Uszkoreit. (1997). *An annotation scheme for free word order languages. In Proceedings of the Fifth Conference on Applied Natural Language Processing* (ANLP), Morgan Kaufmann, San Francisco.

Tanenhaus, M. K., & Brown-Schmidt, S. (2008). Language processing in the natural world. In B. C. M. Moore, L. K. Tyler, & W. D. Marslen-Wilson (Eds.), The perception of speech: From sound to meaning. *Transactions of the Royal Society B: Biological Sciences*, 363, 1105–1122

Thompson-Schill, S. L., Ramscar, M., & Chrysikou, M. (2009). Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Science* 8(5), 259–263.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Tomasello, M. (2008). *Origins of human communication*. Cambridge, MA: MIT Press

Toulmin, S. (1972) *Human Understanding: The Collective Use and Evolution of Concepts*. Princeton: Princeton University Press.

Tulving, E., & Thomson, D.M., (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352-373

Twain, M (1880) [S. Clemens] *A Tramp Abroad.* Hartford, CT: American Publishing Company.

Van Son, R., & Pols, L. (2003). How efficient is speech? *Proceedings of the Institute of Phonetic Sciences*, 25,

Vigliocco, G., T. Antonini and M. F. Garrett (1997). Grammatical Gender Is on the Tip of Italian Tongues. *Psychological Science*, 8/4, 314-317

Van Son, R., & Pols, L. (2003). How efficient is speech? Proceedings of the Institute of Phonetic Sciences, 25, 171–184

Weaver W., and Shannon, C.E., (1963). *The Mathematical Theory of Communication*. Univ. of Illinois Press

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford, England: Blackwell

Yang, C. (2006) *The Infinite Gift*. New York: Scribner's.

Zipf, G.K. (1935). *The Psychobiology of Language*. New York: Houghton-Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least-effort*. Cambridge, MA: Addison-Wesley.

Zubin, D & Köpcke, K-M (2009). Gender Control - lexical or conceptual? In: Patrick O. Steinkrüger/Manfred Krifka (Hrsg.): Trends in Linguistics. On Inflection. Berlin: de Gruyter, pp. 237-262.

Zubin, D & Köpcke, K-M (1996). Prinzipien für die Genuszuweisung im Deutschen.In: Ewald Lang und Gisela Zifonun (Hrsg.): *Deutsch typologisch. Jahrbuch des Instituts für Deutsche Sprache 1995*. Berlin: de Gruyter, 473–491.

Zubin, D & Köpcke, K-M (1986). Gender and Folk Taxonomy: The Indexical Relation Between Grammatical and Lexical Categorization. In: C. Craig (Hrsg*.): Noun Classification and Categorization.* (Typological Studies in Language; Vol. 7). Philadelphia: Benjamins, North America, pp. 139–180.