Linguistic structure emerges from cognitive mechanisms

Molly L. Lewis

Department of Psychology, Stanford University

Conceptual Analysis of Dissertation Area

6 October 2014

Advisor: Michael C. Frank

Additional Readers: Ellen Markman and Noah Goodman

Abstract

Language is highly structured — from the way meaning is organized into categories to elaborate regularities in grammar. Where does this structure come from? In this paper, I argue that linguistic structure is causally related to linguistic usage. In Part I, I begin by arguing that language use is a particular kind of coordination problem. I suggest that Horn's taxonomy of pragmatic pressures (1984) provides a useful framework for understanding the nature of this coordination problem, in the case of language use. I then highlight a number of phenomena in linguistic use that are reflected in linguistic structure. In Part II, I propose a causal process underlying the similarity between usage and structure. This analysis relies on a division of language into five different timescales: pragmatic, discourse, developmental, cultural, and evolution. I conclude by surveying a range of cognitive phenomena that emerge from the dynamics between these timescales.

Human society can be viewed as a field which both influences the individual members of the group and is influenced by them.

– G.K. Zipf, *1949*

Introduction

"Room for cream?" asked the barista. "Mm, yes – just a bit" replied the customer. Mundane linguistic interactions such as this are the building blocks of daily experience. They are individuals making sounds to each other in an effort to coordinate their behavior in the physical world (H. H. Clark, 2006). These interactions are messy, variable, and highly unconstrained. Indeed it is this variability that gives language its vast expressive power (Hockett, 1960). Yet, despite this appearance of irregularity, rich patterns in linguistic usage are revealed when we aggregate across instances of language use both within and across languages. At the level of syntax, for example, there is a strong bias in English to put subjects before verbs and, across languages, this pattern is attested more often than would be expected by chance alone (Dryer, 2005). These types of probabilistic regularities exist at every level of linguistic structure — from phonology, to semantics, syntax, and discourse — and researchers from a variety of disciplines have taken as their project the goal of characterizing these regularities.

In this paper, I argue that we can gain insight into the character of linguistic structure by considering the dynamics of language use. I will suggest the best way to do this is by framing language use as an instance of a broader phenomenon: social interaction (H. H. Clark, 1996). In particular, I will adopt the formal framework of social interaction proposed by Schelling (1980) in which social interactions are viewed as acts of solving coordination problems. To illustrate, consider the barista example above. In this example, the agents are the barista and the customer, and they must coordinate how to fill the coffee mug. There are two outcomes — full and almost

full — and the barista's desired outcome is determined by the preference of the customer. In this case, the barista and the customer rely on language to coordinate their behavior, but this coordination could have been achieved in other ways (e.g. the customer could have shook her head, pointed to the place inside the mug that she wanted the coffee filled to, etc.). Coordination of their behavior is achieved by arriving at the mutually preferred outcome (the customer's mug is almost full).

A key tenet to the broader argument is that the act of using language is itself an act of solving a coordination problem (H. H. Clark, 1996). When a person speaks, there are many possible ways the utterance could be interpreted, and arriving at the intended interpretation is an act of coordination with the listener. For example, in the case of the customer's interaction with the barista, there are many possible interpretations of the phrase, "Room for cream?." The barista could mean "Would you like to add cream to your coffee? If so, I will facilitate that by not filling your mug full with coffee." Or, "We have so much extra inventory of cream! Do you have room in your bag to take some?" Or, "Do you like the band 'Room for cream'?". Or, if the speaker is speaking another language, a totally unrelated meaning. The point is that the speaker's intended meaning is underspecified from the language form alone and the interlocutors must work collaboratively to arrive at a shared understanding. Following D. Lewis (1969), I will suggest that we can gain insight into the dynamics of linguistic coordination problems by using Schelling's formal framework. This perspective on language use will ultimately provide a helpful framework for understanding the relationship between language use and language structure.

It is worth reflecting on the historical relationship between these two aspects of language. Across many schools of linguistics, theorists have made a theoretical cut between language use and language structure: *parole* vs. *langue* (Saussure, 1916), *token* vs. *type* (Peirce, 1931), and *performance* vs. *competence* (Chomsky, 1965). These theorists have different views on the

ontological status of structure — Saussure suggests it is a social fact, while Chomsky argues it is fundamentally a cognitive phenomenon — but they nonetheless agree that there is some sort of invariance in language and it should be the focus of study. Language use has often been seen as an irregular, variant, and epiphenomenal to the true subject of study: structure. However, a number of more recent movements have begun to focus on language use. Labov's (1972) work was an important challenge to exclusionary focus on abstract structure. His work revealed systematicity in the variation of phonology as function of social variables, suggesting that "messy" language use was governed by regularities and could therefore be studied scientifically. The study of pragmatics, more generally, can be seen as a step to find regularity in language use. The goal of this paper is to suggest that, not only are these two aspects of language deeply related to each other, but that the key to understanding linguistic structure may lie in understanding linguistic use.

In Part I, I will outline the linguistic coordination problem as a paradigmatic case of the social coordination problem. I will suggest that linguistic coordination problems are solved through the dynamics of two opposing two forces — the goals of the speaker and the hearer. Following Lewis, I suggest that these opposing forces are resolved by finding an equilibrium point. I will then argue that the equilibria that are reached in language use are reflected in the structure of language, and survey a variety of phenomena in linguistic structure that show this pattern.

In Part II, I will consider the mechanism that might cause linguistic structure to reflect the equilibria reached in linguistic use. I describe five theoretically distinct timescales associated with language, and argue that dynamics between adjacent timescales is responsible for the ultimate emergence of linguistic structure. Given this framework, I will describe a variety of cognitive phenomena that result from the dynamics of these timescales.

Part I: Linguistic structure reflects pragmatic equilibria

Where does linguistic structure come from? Christiansen and Chater (2008, 2010) propose a compelling theory. They argue that multiple cognitive constraints dynamically influence language evolution. They suggest four constraints: the representational format of thought, properties of the percepto-motor system, learning and processing constraints, and constraints that result from reasoning about others' intentions (*pragmatic* constraints). Their argument is that these constraints influence language at the moment of use, but over time, these biases become instantiated in the structure of language. Although each of these constraints likely plays an important role in the evolution of language, the present paper focuses on the independent contribution of pragmatic constraints. The claim is that pragmatic constraints that play out at the moment of language use become fossilized in the structure of language over time. To develop this claim, we begin by modeling language use as a type of social coordination. We then turn to an analysis of language use as a social coordination problem. Finally, we consider three cases where there are similarities in phenomena between language use and language structure.

Social interaction as a coordination problem

Many theorists of language (Zipf, 1936; D. Lewis, 1969; Grice, 1975; H. H. Clark, 1996) have observed that language is an instance of a much broader class of behavioral phenomena — social coordination. For each, language use is a case of multiple agents making interdependent rational choices. By adopting work from game theory, D. Lewis (1969) formalized the notion of language as a coordination problem. He defines a *coordination problem* as follows:

> Two or more agents must each choose one of several alternative actions. Often all the agents have the same set of alternative actions, but that is not necessary. The outcomes the agents want to produce or prevent are determined jointly by the actions

		customer	
		full	almost full
barista	full	0,0	0,0
	almost full	0,0	1,1

Table 1

A payoff matrix of a simple interaction — a barista trying to determine how full to fill a customers' cup. Given the customer's preference for cream, there lies an equilibrium point at 'almost full' for both the barista and the customer. To arrive at this equilibrium point, the two must coordinate.

of all the agents (p. 8).

The key feature of these problems is that some combinations of the agents' choices are better than others: there are a set of joint choices in which no agent would have a larger payoff had the agent alone changed her choice. We refer to these as *equilibrium points*.

This broad framing can describe the dynamics of many social interactions. Take the above case of the barista and the customer, for example. We can model this interaction using a payoff matrix (Table 1). In the matrix, we represent the customer's payoff along the rows and the barista's payoff along the columns. There are two possible choices for level of coffee in the cup—full and almost full — and so each agent gets two rows or columns. The agents relative payoffs are indicated in the cells, with the barista's on the left, and the customer's on the right. This happens to be a very simple equilibria — there is one, and only one, possible set of actions in which is an equilibrium. The customer prefers almost full and the barista fills the cup to almost full. The problem is that the barista does not know a priori where this equilibrium lies, i.e. that the customer's pay off for almost full is 1, relative to 0 for full. To solve this coordination problem,

		Group A	
		food	alcohol
Group B	food	0,0	1,1
	alcohol	1,1	0,0

Table 2

A payoff matrix for a social interaction in which there are two equilibria. Neither group cares who brings what commodity on the vacation; they only care that they bring different things.

the customer and barista make use of language.

More complicated coordination problems arise when there are multiple possible equilibria. Consider a weekend trip in which food and alcohol must be brought. To distribute the burden, half of the vacationers will bring food and the other half alcohol. In this case, the payoff matrix might look something like Table 2. Neither group — Group A or B — has a strong preference about which of the two commodities each brings. However, what is important is that one group brings food and the other alcohol (no one will be happy on a weekend trip with only food or only alcohol). There are thus two equilibria, one at each set of choices where the two groups bring different things. By chance, the vacationers are equally likely to end up in a non-equilibrium as they are an equilibria. They must therefore coordinate, via language or some other means, to ensure that they end up at an equilibrium.

What are the psychological forces that support these coordination games? Zipf's theory of human behavior (1949) provides insight. He argued that all human behavior could be accounted for by a single principle: people are motivated to minimize effort (*The Principle of Least Effort*). To understand this principle, consider a context in which an individual needs to exert some

physical effort, say in walking to a park. The principle predicts they should be motivated to find the solution that minimizes how much effort required (i.e., by finding the shortest path). Critically, Zipf argued this simple principle had explanatory power at the level of social groups. He claimed that this principle operates at the level of the individual, but with interaction, this principle leads to an equilibrium in behavior at the level of the group.

This simple theory of behavior provides a parsimonious account of a wide range of social phenomena. A particularly clear example is the organization of people into social groups in physical space (Zipf, 1949).[1] Zipf assumes that every individual in a society is both a consumer and producer of goods. Governed by the Principle of Least Effort, the individual should minimize effort in terms of movement across land, by consuming (i.e. living) and producing (i.e. working) at the same location. However, as the number of raw goods increases this becomes increasingly difficult because the consumer cannot not live at the doorstop of every finishing plant. This creates a conflict. On the one hand, there is a tendency to diversify, so that the population lives at the doorstep of the production line. This creates a pressure for many physically separated communities, each producing a single good, but with little trade between communities. On the other hand, there is force to unify so that it is easier to trade final goods. The net result is an equilibrium where people live in many different urban centers across the land. This general theory is reflected in more modern theories of urban spatial layouts (Mills, 1967; Brueckner, 1987).

Importantly, Zipf's principle does not provide a description of how individuals come to solve pure coordination problems like the vacationer example above. In that case, it is not clear how two individuals with the exact same payoff structure would arrive at the same, otherwise

[1] A second, well-studied example is economics. At the level of the individual, the consumer tries to minimize something of value, but in this case the valued commodity is money, rather than physical energy (though these are arguably related in important ways). This is the study of microeconomics. With interaction among consumers, this single force leads to regularities at the level of the social group, or the economy— the study of macroeconomics.

arbitrary solution. To address this issue, we need the idea of *convention* which will return to in Part II. However, Zipf's principle does provide insight into how a single psychological force, shared by all individuals, can lead to biases for different alternatives (e.g., a preference for the shortest path to a location). When individuals with these same biases interact with each other, we see the emergence of an equilibrium.

Language use as a coordination problem

The coordination framework can be straight-forwardly applied to language use (D. Lewis, 1969). Language use is a paradigmatic case of a coordination problem because it is a tool that is universal in a community, easy to use, and capable of expressing complex ideas. In the case of language, the core of the coordination problem lies in the resolution of reference. Broadly, resolving reference requires interpreting a meaning from some utterance in a particular context. At the level of individual lexical items, this is a difficult problem because the relationship between linguistic form and meaning is arbitrary (Saussure, 1916; Hockett, 1960). That is, knowing the form of a word does not give a listener any insight into the meaning of that word. Consequently, speakers must coordinate their behavior in order to successfully refer.

The arbitrariness of linguistic form leads to a formal equivalence between the problem of reference and the problems of coordination described above. To understand this similarity, consider a case where there are two novel words, "dax" and "fep," and two novel objects, Object A and Object B. Given this information alone, the listener has no a priori insight into which object each word refers to. This is a problem because the two interlocutors must somehow arrive at the same mappings between words and referents in order to communicate (a system in which you call Object A "fep" and I call it "dax" is a terrible communication system). The interlocutors must therefore coordinate.

The payoff structure for this problem is identical to the vacationer example above (Table 3).

		Speaker	
		Object A–"dax" Object B–"fep"	Object A–"fep" Object B–"dax"
Listener	Object A–"fep" Object B–"dax"	0,0	1,1
	Object A–"dax" Object B–"fep"	1,1	0,0

Table 3

A payoff matrix of the mapping problem. Given two words and two referents, the mappings are arbitrary. The only constraint is that no word should map to more than one object, and no object should map to more than one word. Thus, as in the vacationer example, speakers must coordinate.

Because language is arbitrary, neither speaker cares whether you call Object A "fep" or "dax;" they only care that their mappings are the same. These general dynamics are true not only of individual lexical items, but of all cases of reference. Consider again our example of the barista and the customer. In interpreting the phrase, "Room for cream?," the individual lexical items are relatively unambiguous — presumably both know what "room" and "cream" mean. But, the intended meaning of the entire phrase is underspecified, and so the interlocutors must work together to resolve meaning.

In this framework, we can think of *pragmatics* as the study of the psychological processes that lead to an equilibrium in these referential coordination problems. Pragmatics, then, is just a specific case of the dynamics described by Zipf (1949). In the case of language, Zipf's insight was that speech could be thought of as its own economy, similar to any other social system.

Speech has a physical cost and could be used as tool. He suggested the speaker and the listener were both governed by the Principle of Least Effort and this lead to an equilibrium. In the case of the speaker, effort could be minimized if there existed a single word w that could be used to refer to the set of all concepts C. In the case of the listener, effort would be minimized (in terms of understanding) if there existed a unique word for each unique concept c. The dynamics of the interaction of these two opposing forces is pragmatics. Many theorists have tried to account for pragmatic regularities in behavior, most notably Grice (1975). However, Horn (1984) presents a particularly parsimonious theory that closely aligns with Zipf's more general formulation. He posited two principles that describe the manifestation of Zipf's Principle of Least Effort for the speaker and the hearer each.

> SPEAKER: Say no more than you must. (*Principle of Necessity*)
>
> HEARER: Say as much as you can. (*Principle of Sufficiency*)[2]

These principles are quite straight forward. As a speaker with a necessary meaning to contribute, you want to say a little as possible, while still conveying the intended meaning. In contrast, as a listener, you want the speaker to say as much as possible to minimize your effort at arriving at the correct interpretation. That is, you want the speaker to use sufficient language. In the limit, each strategy on its own does not result in a successful communication system. This is because the speaker's and the hearer's goal are fundamentally linked: While in the short term, in might be less effort for the speaker to utter a single sound, "blah," to convey her meaning, this will lead to confusion on the part of the hearer, which the speaker will then have to clarify with additional language. The speaker and hearer must therefore resolve their two opposing forces in what Horn called a "division of pragmatic labor" in order to arrive at an equilibrium point (Horn, 1984, p.

[2]Horn refers to these as the R and Q principles, respectively. I've opted for less opaque terminology.

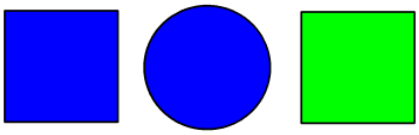

Figure 1. Example stimuli from Frank and Goodman (2012). In this case, the intended referent is the middle shape. Both speaker and hearer agree that the utterance "circle" is an equilibrium point, as compared to "blue," even though both terms equally describe the intended referent in truth-functional terms.

22).[3]

While these principles are simple, the dynamics they give rise to are complex. Frank and Goodman (2012) present a formal model that captures these dynamics in a simple reference game. In their game, there are three possible referents and each referent has two relevant features (Fig. 1). Given the constraint that a speaker can only utter a single word, there are always two possible words the speaker could utter. For example, if the intended referent is a blue circle, a speaker could use either the word "blue" or "circle" to refer to the object. Critically, both words are equally true of the referent from the perspective of truth functional semantics. The phenomenon that this model captures is that the speakers use different words to refer to an object — and that listeners expect them to — depending on the context in which the word is uttered. In particular, speakers tend to choose a word that most uniquely identifies the intended referent, given the referential context. In other words, they select the word that is most informative. For

[3]Note that there is a super maxim that is also operating here: the cooperative principle (Horn, 1984; Grice, 1975). The cooperative principle is essentially the idea that the interlocutors realize that utterances are the result of an equilibrium from the dynamics of these two forces. Put another way, it is a statement that the interlocutors realize that they are playing a coordination game.

example, in the example trial pictured in Figure 1, speakers tend to use the word "circle" instead of "blue" to identify the middle referent. Frank and Goodman use a Bayesian framework to formalize this notion of informativity, and their model closely captures the behavioral data in this reference game. This suggests that the behavior of both interlocutors is guided by a tacit understanding of informativity in the referential context.

This notion of informativity falls directly out of Horn's principles. In this reference game, speakers are constrained by the length of the utterance they can use (one word), but the choice of words is free. From the listener's perspective, the utterance must be sufficient and so uttering "blue" would be insufficient in the above context because it is ambiguous, and there is a better alternative. From the speaker's perspective, the utterance must be necessary, but should not be too verbose. This force is enforced by structure of the task — the speaker must contribute something to participate in the task, and the utterance cannot be overly verbose given the single word constraint. Maximal informativity — a sufficient contribution, given the context — is thus the equilibrium between these two forces.

Pragmatic equilibria reflected in the structure of language

Simple reference games like that of Frank and Goodman (2012) are one example of the kind of equilibria that emerges from the interaction of Horn's principles, but there are many others. In the present section, we consider how these pragmatic equilibria points in language use are reflected in the structure of language. We will consider this relationship for three different kinds of linguistic structure: semantics, words, and syntax.

Semantics. Semantics concerns the context-independent meaning associated with a word. The size of the semantic space denoted by a particular word reflects an equilibrium point between Horn's speaker and hearer principles. From the hearer's perspective, Horn argues there is a pressure to narrow semantic space (Horn, 1984). This reflects the idea that the hearer's optimal

language is one in which every possible meaning receives its own word. One example of this is the word "rectangle." This word refers to a quadrilateral with four right angles. A special case of a "rectangle" is a case where the four sides are equal in length, which has its own special name, "square." Consequently, the term "rectangle" has been narrowed to mean a quadrilateral with four right angles, where the four sides are *not* equal.[4] From the speaker's perspective, there is a pressure for semantic broadening. This is because the speaker's ideal language is one in which a single word can refer to a wide range of meanings. An example of this is the broadening of brand names to refer to a kind of product. For example, "kleenex" is a name of a product name for facial tissues, but has taken on the meaning of facial tissues more generally.

The opposition of these two semantic forces predicts an equilibrium in the organization of semantic space that satisfies the pressures of both speaker and hearer. A body of empirical work has tested this prediction by examining the organization of particular semantic domains cross-linguistically (Regier, Kemp, & Kay, 2014). Languages show a large degree of similarity in how they partition semantic space for a particular domain, which is likely due to universal cognitive constraints. But, they also show a large degree of variability and these different systems can be shown to all approximate an equilibrium point between speaker and hearer pressures.

Kemp and Regier (2012) demonstrate this systematicity in the semantic domain of kinship. For each language, they developed a metric of the degree to which Horn's speaker and hearer pressures (in their terminology: communicative cost and complexity, respectively) are satisfied. A language that better satisfies the hearer's pressure is one that is more complex, as measured by the length of the description of the system in their representational language. A language that better satisfies the speaker's pressure is one that requires less language to describe the intended referent. To understand this, consider the word "grandmother" in English: this word is ambiguous in

[4]Horn also points out that there are cases of narrowing that are speaker-based, as in "drink" for "alcoholic drink."

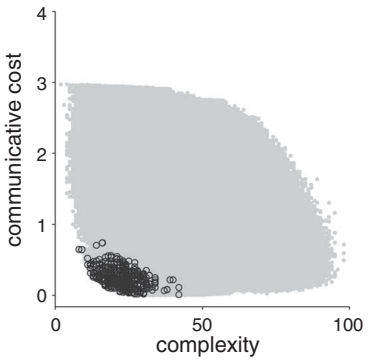

Figure 2. Plot from Kemp and Regier (2012). Languages organize the semantic space of kinship in a way that optimizes both speaker and hearer pressures. The notion of *communicative cost* maps onto Horn's speaker principle and the notion of *complexity* maps onto Horn's hearer principle. Gray circles represent possible systems and black circles represented actually attested languages. The key observation is that all of the attested languages are clustered around the bottom left corner that corresponds to an equilibrium between speaker and hearer pressures.

English because it could refer to either the maternal or paternal mother, and so identifying one in particular is more costly in English than in a language that encodes this distinction lexically. They find that the set of attested languages is a subset of the range of possible languages, and this subset partitions the semantic space in a way that is near the optimal tradeoff between speaker and hearer pressures (Fig. 2). This type of analysis has also been done for the domains of color (Regier, Kay, & Khetarpal, 2007), light (Baddeley & Attewell, 2009), and numerosity (Y. Xu & Regier, 2014).

A second phenomenon that is predicted by these forces is the presence of lexical ambiguity. That is, cases in which there are multiple meanings associated with a word, from a context-independent perspective. Language is rampant with examples. For example, the word "bat" could mean either the instrument used in baseball or the flying mammal. This type of ambiguity in context-independent meaning is tolerated because the meaning is usually easily

disambiguated by context. When the word "bat" is uttered while watching a baseball game, the mammal usage of the word is very unlikely. We can view the presence of this ambiguity as an equilibrium in Horn's speaker and hearer principles. If the meaning of a word can be disambiguated by the referential context, then it would violate the speaker's principle to have an overly-specific term for a meaning.

Indeed, recent work by Piantadosi, Tily, and Gibson (2012) reveals systematicity in the presence of lexical ambiguity in language. They argue that ambiguity results from a speaker based pressure to broaden the meaning of a word to include multiple possible meanings. In particular, they suggest that this pressure should lead to a systematic relationship between the presence of ambiguity and the cost of a word. According to their argument, costly words (in terms of length, frequency, or any metric of cost) that are easily understood by context violate the speaker's principle to say no more than you must. Consequently, there should be a pressure for these meanings to get mapped on to a different, less costly word. This word may happen to already have a meaning associated with it, and so the result is multiple meanings being mapped to a single word. For example, in the case of the word "bat," a speaker could instead say "baseball bat." But, because this referent is easily disambiguated in context from the mammalian meaning, Horn's speaker principle leads to a pressure to use the shorter form.[5] This leads to a testable prediction that shorter words should tend to be more ambiguous. Through corpus analyses, Piantadosi et al. (2012) find this precise relationship between cost and ambiguity. They find a linear relationship between word length and ambiguity across English, Dutch and German:

[5]There are also cases of temporary ambiguity in the meanings of syntactic structures. For example, in a sentence that begins "The coach knew you..." it is unclear whether "you" is a direct object or the subject of a relative clause. Speakers can avoid this ambiguity by inserting a "that," but this is optional. Work by Ferreira and Dell (2000) suggests that speakers often do not avoid this ambiguity, presumably because the intended meaning can easily be recovered from context.

Shorter words are more likely to have multiple meanings.

An additional case of this lexical ambiguity is found in words that have very little context-independent meaning, known as indexicals or deictics (Frawley, 2003). These words get their meaning from the particular referential context of the utterance, and are therefore highly ambiguous from a context-independent perspective. There are many types of indexicals that are present to varying degrees across languages. An example of a temporal indexical form is "tomorrow." The context-independent meaning of this word is something like "the day after the day this word is being uttered in." Critically, abstracted from any context, this word has little meaning; it is impossible to interpret without having knowledge about the day the word was uttered. This phenomenon is also present in person pronouns (e.g. "you" and "I") and spatial forms, like "here" and "there." As for lexical ambiguity, this type of ambiguity is a predicted equilibrium point from Horn's principles: If the hearer can recover the intended referent from context, the speaker would be saying more than is necessary by using an overly-specific referential term (e.g., "December 18th, 2014" vs."tomorrow"). Language structure reflects this pressure through lexicalized ambiguity in the form of indexicals.

Finally, the relationship between the meanings of different words can be seen as a consequence of Horn's principles. A number of theorists have noted a bias against two words mapping onto the same meaning — that is, a bias against synonymy (Saussure, 1916; Kiparsky, 1983; Horn, 1984; E. Clark, 1987, 1988). This bias is an equilibrium between Horn's speaker and hearer principles. Recall that the optimal language for a hearer is one in which each meaning maps to its own word — exactly a language biased against synonymy (see Fig. 3). It turns out that the speaker's pressure also biases against synonymy. The optimal language for the speaker is a language where a single word maps to all meanings. But, a case where multiple words map to a single meaning is also undesirable because the speaker must keep track of two words. So, for both

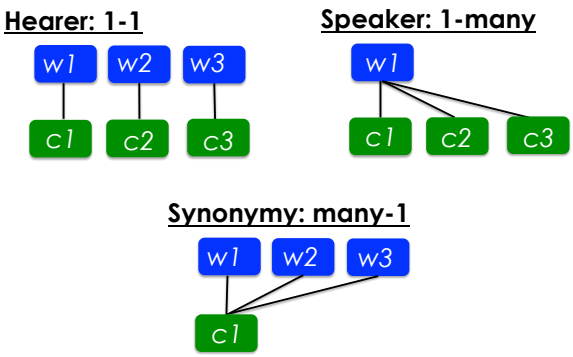

Figure 3. Three possible structures on the organization of lexicon. According to Horn's principles, the hearer's optimal language is one in which there is a one-to-one mapping between words and concepts. The speaker's optimal language is one in which there is one word that maps to many concepts. Synonymy, or a many-1 structure, is thus dispreferred by both interlocutors.

the speaker and the hearer, there is pressure to avoid synonymy. Thus, when a listener hears a speaker use a second word for an existing meaning, the hearer infers that this could not be what the speaker intended because this would violate the speaker's principle. The result is an assumption that the second word maps to a different meaning. This pattern is reflected in language structure by a one-to-one pattern in the lexicon — that is, a structure in which each word maps to exactly one meaning and each meaning maps to exactly one word.

As one kind of evidence for this one-to-one structure in the lexicon, Horn (1984) points to a phenomenon called *blocking*. Blocking refers to cases in which an existing lexical form blocks the presence of a different, derived form with the same root. Consider the following examples:

(a) fury furious *furiosity

(b) *cury curious curiosity

In both (a) and (b), forms that would be expected, given the inflectional morphology in English, are not permitted. This is presumably because they would have the same meaning as the existing

form because they have the same root. Examples such as this provide some evidence for a one-to-one structure in language, but a one-to-one structure is a particularly difficult linguistic regularity to test empirically. Nonetheless, it is an important regularity because it licenses certain inferences in interpreting the meaning of words. In particular, the cognitive representation of a one-to-one regularity has been posited as an explanation of children's bias to map a novel word onto a novel object (Markman & Wachtel, 1988; Markman, Wasow, & Hansen, 2003). We return to this issue in Part II.

Words. Horn's principles make a prediction about the relationship between the length of utterances and their meanings. In many cases, it is possible to use two different utterances to refer to the same meaning (in truth functional terms), and often these utterances differ in length. Horn (1984) presents the following example:

(1a) Lee stopped the car.

(1b) Lee got the car to stop.

Both (a) and (b) have the same denotational meaning (the successful stopping of a car), but they differ in length ((b) has two extra words). Horn argues that this asymmetry leads to an inference on the part of the listener that the two differ in meaning. The logic of this inference is identical to the lexical structure case above. The listener hears a speaker use a more costly phrase to express a meaning that could have been expressed in a less costly way. The listener thus infers that this other meaning could not be what the speaker intended because this would violate the speaker's principle to say no more than is necessary. Horn adds an additional layer to this argument. He suggests that no only do these two forms differ in meaning, but that they map onto meanings in a systematic way. In particular, he argues that the longer form gets mapped on to the more marked meaning, while the shorter form refers to the unmarked meaning. The notion of 'markedness' is underspecified here, but an intuitive definition is related to complexity: more marked things are

		Speaker	
		{short–simple, long–complex}	{short–complex, long–simple}
Listener	{short–complex, long–simple}	0,0	1,1
	{short–simple, long–complex}	1,1	0,0

Table 4

The payoff matrix for the speaker and listener in solving a coordination problem in which there are two words — one short and one long — and two meanings — one simple and one complex. Given these constraints, there are two equilibrium lexicons. As observed by Horn (1984), speakers tend to arrive at the equilibrium in the bottom left corner.

more conceptually complex, while less marked things are more conceptual simple. Thus, in the above example, (a) would refer to a simple, average case of car stopping, while (b) might refer to case where something complex or unusual happened, perhaps because Lee used the emergency brake.

The source of the particular mapping between forms of different lengths and meanings of different degrees of markedness is unclear. This is because, in principle, there are multiple equilibrium points in the mapping between form and meaning. Assuming a one-to-one constraint on the mapping, there are two possible equilibria: {short–simple, long–complex} or {short–complex, long–simple} (Table 4). Both satisfy the constraint that each form gets mapped to a unique meaning. So how do speakers arrive at the {short–simple, long–complex}

equilibrium? This is a difficult result to derive from models of pragmatic reasoning. Bergen, Levy, and Goodman (in prep) successfully derive this result as a consequence of the fact that {short–simple, long–complex} is a more optimal mapping for the speaker (the indirect result of Zipf's Principle of Least Effort). Another possibility relies on iconicity: hearers have a cognitive bias to map more complex sounding forms to meanings that are similarly complex.

Despite the absence of clear theoretical account of this phenomenon, the empirical data suggest that learners do indeed arrive at the predicted equilibrium. Bergen, Goodman, and Levy (2012) provide evidence for this type of implicature in a communication game. In their task, partners were told that they were in an alien world with three objects and three possible utterances of different monetary costs. They operationalize the idea of markedness or complexity as frequency, such that participants were instructed that each of the three different objects had three different base rate frequencies associated with them. Participants' task was to communicate about one of the objects using one of the available utterances. If they successfully communicated, they received a reward. The results suggest that both the speaker and hearer expected costlier forms to refer to less frequent meanings. This study provides one data point suggesting that Horn's predicted equilibrium between word length and meaning emerges in coordination games.

There is a growing body of evidence suggesting this equilibrium is also reflected in the structure of words. One approach to testing this hypothesis is to use the linguistic context of a word to measure the complexity of meaning. The idea is that words that are highly predictable, given the linguistic context, have more complex meanings, while words that are less predictable given the linguistic context, have less complex meanings. Piantadosi, Tily, and Gibson (2011) measured the relationship between the predictability of words in context and the length of words. Across 10 languages, these two measures were highly correlated: words that were longer were less predictable in their linguistic context on average. This result held true even controlling for the

frequency of words. Additional evidence for this relationship comes from examining pairs of words that have very similar meaning, but differ in length (e.g. “exam” vs. “examination;” Mahowald, Fedorenko, Piantadosi, & Gibson, 2012). Through corpus analyses, they find that the longer forms are used in less predicable linguistic contexts. In a behavioral experiment, they also find that speakers are more likely to select the longer word in unsupportive contexts. This body of work points to a systematic relationship between word length and meaning, when complexity is operationalized as predictability in the linguistic context.

Some of our own work provides more direct evidence for this equilibrium. Given a novel word, we find that both adults and preschoolers are more likely to map a longer word to a more complex object, as compared to a short word (M. Lewis, Sugarman, & Frank, 2014). A key difference between our work from prior work is that we directly manipulate the complexity of word meaning, rather than using the predictability of linguistic context as a proxy. We have operationalized complexity in three different ways. The first is to directly manipulate the number of object parts the referent has. Second, we have measured complexity by obtaining complexity norms from participants on real objects. Third, we have operationalized complexity through a reaction time measure. In each case, we see a bias to map longer words to more complex referents, as compared to a short word. We also find this bias in natural language. We asked participants to rate the complexity of the meaning of 499 English words, and found that these ratings were highly correlated with word length in both English and 79 other languages. Taken together, this work provides strong evidence that the equilibrium between word length and complexity of meaning found in coordination games, such as Bergen et al. (2012), is also reflected in the structure of the lexicon.

Syntax. The order of words, or syntax, is another level of language structure that reflects equilibria of language use. Levinson (2000) provides a detailed account of how Chomsky’s

syntactic binding constraints can be reinterpreted as pragmatic equilibria. Chomsky argues there are three different principles that govern how pronouns can be co-indexed with their antecedents. I will highlight two here. The first principle is that an anaphor (like "herself") is bound in its governing category (e.g., a sentence). The second is that a pronoun cannot be co-indexed with a c-commanding[6] noun phrase. This constraint provides an of account why (a) below is grammatical, but (b) is not.

(a) $Elsa_1$ *likes* $herself_1$.

(b) *$Elsa_1$ *likes* her_1.

In (a), the pronoun "herself" is an anaphor and thus is grammatical by Chomsky's first principle. In contrast, in (b), "her" is co-indexed with a c-commanding noun phrase ("Elsa"), and thus is ungrammatical. Levinson offers an alternative explanation based on pragmatics. The explanation relies on the same logic that leads speakers to avoid synonymy. The logic can be informally summarized as follows:

1. "herself" is an anaphor and is therefore co-indexed with a noun phrase in the local domain. (identical to Chomsky's first principle, but could also be motivated pragmatically)
2. In (b), the speaker used a different form ("her").
3. If the speaker had meant to refer to the antecedent in the local domain, she would have used "herself" because it is more informative. (by the hearer principle)
4. The speaker did not, and thus the intended interpretation must be an antecedent outside the local domain. (i.e., $Elsa_1$ *likes* her_2.)

[6]The term *c-command* refers to a specific relationship in generative grammar which is not relevant here. Broadly, it suffices to say that subjects c-command objects in English.

This account, which relies only on general principles of pragmatics, is able to account for the observed pattern of grammaticality judgments.

Several experimental findings also suggest that there are pragmatic equilibria reflected in syntax. One of the primary patterns of linguistic structure to be explained is the linear structure of words. In particular, why some word orders are much more prevalent across languages than others. Gibson et al. (2013) offer an account of one aspect of this regularity and this account can be interpreted as a suggestion that a pragmatic equilibrium is reflected in the structure of language.

Their argument is as follows. There is some evidence that subject-object-verb (SOV) word order is the cognitive default (e.g. Senghas, Kita, & Özyürek, 2004). This proposal reflects the fact that SOV order is the most prevalent word order cross-linguistically (47%). But a puzzle still remains: If there is a SOV bias, why is a second order — SVO — almost equally as prevalent (41%)? Gibson et al. (2013) propose that the move from an SOV to SVO word order reflects a communicative pressure. The idea is that subjects and objects are often semantically confusable (because they are both entities), and thus, it is possible in the "noisy channel" of communication for one of the noun phrases to get deleted. This leads to confusion on the part of the hearer. Thus, they argue that the two should be linearly separated with the verb argument in order to avoid confusion. This is advantageous because if an argument is deleted, the correct grammatical relation of the communicated argument can be inferred from the input in an SVO order, but not an SOV order.

Though they motivate this prediction from an information theoretic perspective, this result can also be derived from Horn's principles. Both the hearer and the speaker want to minimize effort. If it is indeed the case that SOV order frequently leads to confusion on the part of the hearer, this becomes problematic for the speaker who will have to often expend extra effort to

clarify the confusion. Thus, given that there is an alternative word order that requires the same amount of effort — SVO order — there should be a pressure on the part of the speaker to move towards using this order.

To test this proposal, Gibson et al. (2013) asked speakers to produce sentences describing scenes. A critical prediction of their argument is that confusion should be more likely when both arguments are animate (e.g. "The girl pushed the boy") compared to a case when there is an asymmetry in animacy (e.g. "The girl pushed the car"). In their key experiment, speakers of Japanese and Korean viewed brief videos of events. Japanese and Korean speakers were selected because their native languages were not SVO. Following each event, participants used non-verbal gesture to represent the events. Critically, the events involved three nouns that required using and embedded clause structure to describe (e.g., "The woman says the fireman kicked the girl."). The patient of the embedded clause was either animate or inanimate. The measure of interest was the location of the embedded clause as a function of the confusability of the noun phrases. If participants gesture in accord with their native word order, they should show a SOV (S_1 $[S_2$ O_2 $V_2]$ V_1) pattern in their gestures. However, if this bias is sensitive to the confusability of referents, they should be more likely to show a SVO pattern (S_1 V_1 $[S_2$ O_2 $V_2]$) when all the noun phrases are animate. Consistent with the prediction, this is exactly what they found: Japanese and Korean speakers were more likely to use the SVO pattern when the entities were all animate, as compared to when the embedded patient was inanimate.

Case marking is a second case in which pragmatic equilibria are reflected in syntax. In many languages, case marking is used as a syntactic strategy to indicate grammatical relations, like subject or object. Case markers are affixes that attach to the noun root, and are typically used in languages where word order is not a cue to grammatical relations. Case marking is usually optional but principled in its usage based on semantic properties of the nouns like animacy. For

example, speakers often case mark nouns that appear in unpredictable roles, such as when an inanimate noun functions as a subject. Following logic similar to Gibson et al. (2013), Fedzechkina, Jaeger, and Newport (2012) used an artificial learning paradigm to test whether speakers regularize a language to minimize confusion between nouns. They predicted that speakers should be more likely to use case marking when the nouns were all animate, and thus confusable. They exposed learners to a verb-final language with flexible constituent order and optional case-marking. Critically, case-marking in the input language was not conditioned on animacy. Consistent with the prediction, they found that learners tended to mark animate objects with an overt case marker more frequently than inanimate objects (because animate objects are confusable with animate subjects). This provides another case in which pragmatic pressures are reflected in regularities in language structure.

Part II: A casual link between linguistic use and linguistic structure

Why does language structure reflect patterns of language use? In the present section, I propose a speculative answer to this question. The answer posits an indirect causal link between language use and language structure: Language use over time leads to regularities in linguistic structure. I will outline this answer by relying on an analysis of language separated by five different timescales. To preview, the hypothesis is that linguistic structure reflects language use as a consequence of local dynamics between adjacent timescales.

A *timescale* is a unit of time over which significant changes in state occur. For example, the timescale of dinner is about an hour, where the significant changes in state are walking to the restaurant, sitting down, ordering food, eating the meal, then dessert, etc. In contrast, the timescale of gaining weight occurs over weeks, where the significant changes in state correspond to appreciable weight changes. Critically, the length of the timescales is determined by the change of interest.

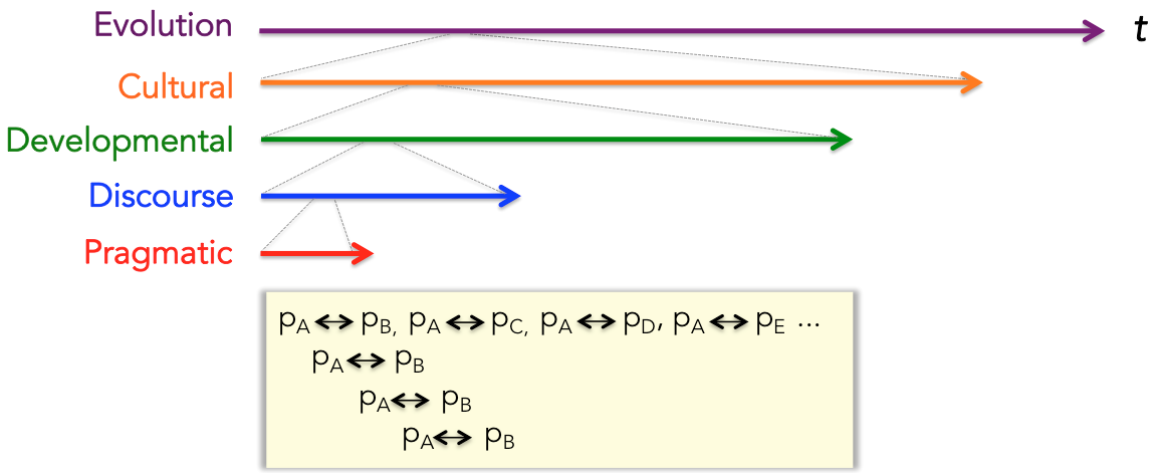

Figure 4. The five linguistic timescales. Each timescale is nested within a slice of longer timescales. The claim is that the dynamics between adjacent timescales (e.g., pragmatic and discourse) lead to changes over time in longer timescales. The developmental timescale is characterized by a particular person from a particular generation, p_A, interacting with a series of people over the lifespan. Repeated interactions with the same person, p_B, characterize the discourse timescale.

In the case of language, there are five timescales over which significant changes occur (Fig. 4). The first is the *pragmatic timescale*. The pragmatic timescale corresponds to the processes described by Horn's principles, and addressed in detail in Part I. The *discourse timescale* is a slightly longer timescale. The discourse timescale corresponds to repeated interactions with the same person; a series of pragmatic interactions. The third is the *developmental timescale*. The developmental timescale corresponds to the lifetime of an individual. It is composed of many interactions (on the pragmatic timescale), some of which with the same people (on the discourse timescale). Many people interacting over their lifetimes lead to change at the *cultural timescale*. This is the timescale over which significant changes in language structure occur (sometimes referred to as the "language change" timescale). Finally, at the longest timescale, is the *evolution*

timescale. All of the dynamics at lower timescales occur within a small slice in evolutionary time. While I will not have much to say about this timescale, its main significance is to situate the present claims with respect to claims about the innateness of language. Following Christiansen and Chater (2008), the suggestion is that there are aspects of language that are innate and constrain the dynamics of shorter timescales. Importantly, however, there are also dynamics that take place at shorter timescales, and these dynamics are the focus of the present section.

The phenomenon to be explained is why dynamics at the pragmatic timescale are reflected in structure at the cultural timescale. The proposal is that there are dynamics between adjacent timescales, and that, over time, these dynamics lead to change on longer timescales. Importantly, the character and phenomena of the dynamics between each pair of timescales are different. For example, the dynamics between discourse and developmental timescales are reflected in cognitive changes in the mind of a particular speaker. In contrast, the dynamics between cultural and evolutionary timescales are reflected in genetic changes in linguistic abilities.

The idea of a causal relationship between language use and structure is not new. One of the earliest proposals of this idea was Whorf (1956) who argued that habitual patterns of talking in particular ways (what he called "fashions of speaking") lead over time to different conceptualizations of the world.[7] Grammar is a case where this view as been particularly well articulated, under the heading of *Emergent Grammar* (Hopper, 1987):

> The notion of Emergent Grammar is meant to suggest that structure, or regularity, comes out of discourse and is shaped by discourse as much as it shapes discourse in an on-going process. [...] Structure, then, in this view is not an overarching set of

[7]This is an important nuance to claims about linguistic relativity that is often over-looked: It is not *that* a language has a label for a concept that matters, but rather the presence of that label in conjunction with a developmental history of using that label.

> abstract principles, but more a question of a spreading of systematicity from individual words, phrases and small sets. (p. 142)

More recently, cognitive psychologists has begun to formally model these dynamics. In this tradition, Bybee and McClelland (2005) write: "Properties of formal structure [...] are facts about the structure that are to be explained as arising from the cumulative impact of the processes that shape each language, as it adapts through the process of language use" (p. 406). They argue for the value of a connectionist framework in capturing these dynamics. Also within a connectionist framework, McMurray, Horst, and Samuelson (2012) highlight the relevance of different timescales in capturing the phenomenon of children's word learning across the developmental timescale. Perhaps the broadest framing of these dynamics has been by Christiansen and Chater (2008), who put propose a mechanism closely aligned with the present argument.

The goal of the present section is to synthesize these many claims about cumulative dynamics into a single framework. In what follows, I describe cognitive phenomena related to the dynamics between each pair of adjacent timescales, beginning with the two shortest timescales: pragmatics and discourse.

Dynamics between pragmatic and discourse timescales

The pragmatic timescale is the locus of language use ("one-shot" communication problems), but regularities emerge when language use is aggregated across multiple interaction with the same speaker. This is the discourse timescale. The dynamics between these timescales are critical to understanding how interlocutors solve the problem of multiple equilibria. Recall coordination problems like the one described in the payoff matrix in Table 3: Speakers must figure out how to map two novel words onto two novel objects. The critical feature of this coordination problem is that there are two equilibrium points. The question then is, how do speakers mutually arrive at the same point? Schelling (1980) argues that speakers may move to

equilibria that are more salient, either perceptually or given prior knowledge. For example, if one of the objects is flashing lights and beeping loudly, while the other is a piece of wood, there might be a perceptual bias to assume that the first word uttered corresponds to the flashing, beeping object. However, if you are helping your partner build a shelf, there might be a bias to select the piece of wood, which is more salient given your knowledge of the speaker. Indeed, there is evidence in the psychological literature that interlocutors make use of salience in solving coordination problems (H. H. Clark, Schreuder, & Buttrick, 1983).

But how do speakers solve the the problem when there is no asymmetry between equilibria? D. Lewis (1969) proposes the notion of *convention* to answer this question. He suggests that once speakers happen to successfully coordinate their behavior at a particular equilibrium, there is inertia to maintain that equilibrium rather than switch to an alternative which is, a priori, equally good. A series of interactions over the pragmatic timescale thus lead to a convention at the discourse timescale. In the case of this example, the idea is that neither "dax" nor "fep" is a better name for Object A. But, once the speakers successfully coordinate by using "dax" to refer to Object A, there is a pressure for both to continue using this linguistic form to refer to Object A. The speakers have thus established a convention and the more this convention is used, the more it becomes entrenched. Knowledge of partner-specific conventions is one aspect of what H. H. Clark (1996) refers to as *common ground*.

In the social psychology literature, there is a large body of work that speaks to the psychological processes that support the emergence of conventions. This work is under the rubric of "conformity," where the idea is that individuals in social groups are motivated to conform to the perceived social norm (Cialdini & Goldstein, 2004). When language use is couched as a particular case of social interaction, this work becomes relevant. The idea is that all mappings between linguistic form and meaning are arbitrary, but individuals are motivated to conform.

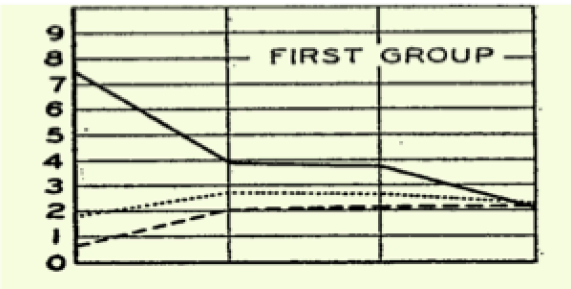

Figure 5. Plot reproduced from Sherif (1935), showing the median distance judgment (in inches) for three different subjects. The *x*-axis corresponds to four different iterations of the experiment: first individually, then three iterations as a group. Over time, the participants converge on an arbitrary norm.

Thus, once an equilibrium is established, speakers conform to the perceived norm (that "dax" maps to Object A, for example). Sherif's autokinetic experiment (1935) provides a powerful demonstration of this pressure. In his task, participants viewed a dot of light on a wall and were asked to indicate when the dot moved and then estimate its distance. In reality, the light never moved. Nonetheless, all individuals reported seeing some movement in the light. Critically, in one version of the study, three strangers were tested individually in the task. They were then tested as a group three additional times. The group context was identical to the individual context except that the subjects could overhear other subjects' responses. Figure 5 plots the median distance judgments of one group of three subjects over iterations of the experiment. When tested individually, the three subjects were highly variable in their distance judgments. However, when tested together, their judgments tended to converge, and this convergence increased over iterations of the experiment. This result provides a clear demonstration of the minimal conditions necessary for interacting social partners to converge on an arbitrary equilibrium.

In the domain of language, this same phenomenon is observed in the way partners establish reference. Because there are multiple ways of referring to an object, interlocutors must establish

some convention. This happens over the discourse timescale. H. H. Clark and Wilkes-Gibbs (1986) demonstrate this phenomenon in the laboratory. In their task, pairs of naive subjects were randomly assigned to either the role of director or matcher. They were seated across from each other with an opaque wall between them. Each subject had a set of cards with ambiguous images that was identical to their partner's. The director's cards were arranged in a grid, and her task was to direct the matcher to organize her cards in the same way using only verbal instruction. After repeatedly completing this task with the same partner, the directors began to use overall fewer words to describe the cards. For example, in trial one, one director used the phrase "the next one looks like a person who's ice skating, except they're sticking their arms out in front," but by trial six the same director simply used the phrase "the ice skater" to refer to that same card. This suggests that the interlocutors arrive at an equilibrium point about how to refer to the different referents, known as *lexical entrainment*. Brennan and Clark (1996) argue that this shared way of referring to an object reflects a *conceptual pact* between interlocutors about how to conceptualize the referents. Consistent with this view, Metzing and Brennan (2003) show that this phenomenon is partner-specific, suggesting that low level cognitive effects (e.g. memory recency) cannot alone account for the observed coordination. Together, this line of work demonstrates how in-the-moment pragmatic pressures lead to equilibrium in discourse.

In addition to reference, there are a number of findings that suggest that interlocutors coordinate other aspects of language. For example, in a study by Branigan, Pickering, and Cleland (2000), interlocutors were found to coordinate their syntactic structures. The task involved completing a picture description task with a confederate in which the two alternated speaking. Critically, sometimes the confederate used a prepositional object structure (e.g. "The girl is throwing the ball to the dog.") and sometimes she used a double-object construction (e.g. "The girl is throwing the dog the ball."). Subjects tended to use the syntactic structure used by the

confederate to describe their own picture (even though there was no semantic overlap between the two), suggesting a coordination of syntactic structure. Other work suggests low level perceptual coordination. For example, Trude and Brown-Schmidt (2012) find that speakers acquire speaker-specific knowledge about how individuals pronounce different words. While the cognitive mechanisms supporting these different types of coordination may differ, each is a case of discourse-level coordination in language use.

Dynamics between discourse and developmental timescales

Many repeated interactions on the pragmatic and discourse timescales lead to change on the developmental timescale. One way to think about this change is as *learning*. This learning is a sort of "cached equilibrium" that results from aggregating over interactions with social partners that each arrives at a similar equilibrium. For example, in the case of semantics, learners make many inferences about the meaning of particular words in many particular interactions across time. The intended referent in each of these contexts is the result of an equilibrium in a coordination problem. Over time, the learner need not reason through the pragmatic logic for each context (e.g., "If the speaker had meant object A, should would have said 'dax' but she didn't, and so...", etc.). Rather she can make use of stored knowledge: speakers tend to use a particular label to refer to things that have roughly these features. The induction of which features map on to a particular label is the process of inducing the semantic structures of language, like those described by Kemp and Regier (2012).

A large body of work documents learners' ability to use of pragmatic information to learn the meaning of new words (e.g. Baldwin, 1991; E. Clark, 1987; F. Xu & Tenenbaum, 2007; Frank & Goodman, 2014), although the developmental trajectory of these skills is not well-agreed upon. For example, Frank and Goodman (2014) used a task analogous to Frank and Goodman (2012), but with novel words. To illustrate their task, imagine someone used the word "dax" rather than

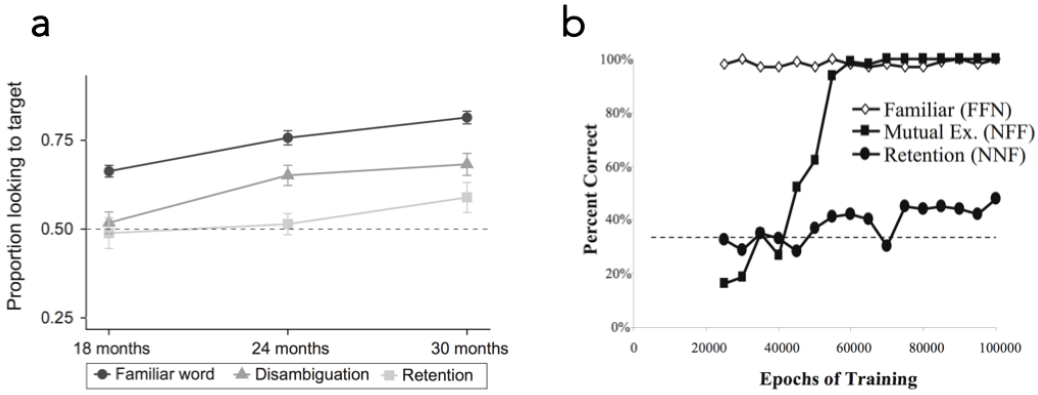

Figure 6. Plots reproduced from (a) Bion et al. (2012) and McMurray et al. (2012). In (a), 24-month-olds show a disassociation between their ability to map a novel word onto a novel object (triangles) and retain that mapping when tested later on (squares). Plot (b) presents a simulation of this result from an associative model. The model is able to successfully map a novel word onto a novel object (squares) before it succeeds in retaining that mapping over time (circles).

"circle" to refer to the middle object in Figure 1. What would you think "dax" meant? As in the familiar word task, the prediction is that learners will assume that speakers are being informative by picking out the smallest set of features that are true of the referent, given the contextual alternatives (in this case, the concept CIRCLE). Preschool age children were found to reason in exactly this way, using the notion of informativeness to guide their inferences about word meaning.

But this in-the-moment inference is not itself learning. Learning, rather, is the result of storing information about these individuals interactions and doing some sort of generalization across them. There is reason to think that in-the-moment inference and learning are two distinct cognitive processes. For example, Bion, Borovsky, and Fernald (2013) tested 24- and 30-month-olds in a task that required children to infer the referent of a novel word in the presence of a familiar and a novel object (often referred to as the "mutual exclusivity" task; Markman &

Wachtel, 1988; Markman et al., 2003). 24-month-olds were able to correctly infer the referent of the word in this context (the novel object), but showed no evidence of remembering this mapping when tested later (Fig. 6a). 30-month-olds, in contrast, both inferred the correct novel word and retained it over a short interval. The fact that these two skills are not coupled in development suggests that they may rely on cognitively distinct processes.

Recent work by McMurray et al. (2012) captures these two distinct timescales in a computational framework. They demonstrate how a wide variety of word learning phenomena can arise from the dynamics of an associative model. They instantiate the role of the in-the-moment pragmatic timescale in terms of the activation of word and object nodes, and the effect of long-term learning in the associative weights on the links between nodes. Of particular note, they are able to capture the empirical pattern observed by Bion et al. (2013, Fig. 6b). In one simulation, they tested a model which knew the meaning of some words but not others. The model was then tested in a setup similar to the Bion et al. (2013) task. In the inference task, the model was presented with two familiar objects, one novel object and a novel word and, with enough training, was able to infer that the novel word mapped to the novel object. A second setup tested retention, in which the model was tested with two novel objects, one familiar object and a novel word. These trials revealed that the model eventually retained the mapping that was made during the inference trials. Critically, however, the rates of the emergence of these patterns differ: the model quickly began to show in-the-moment inference, but only eventually began showing evidence for retention of these mappings. This demonstrates how the discrepant pattern observed by Bion et al. (2013) can emerge from the dynamics of a relatively simple associative model.

In related work, we have tried to understand the psychological forces supporting the bias to map a novel word to a novel object (M. Lewis & Frank, 2013b). There are two broad proposals for explaining this effect in the literature. One proposal is that children rely on pragmatic

reasoning ("Why would you have used that weird word to refer to the familiar object, if you had intended the familiar object?;" E. Clark, 1987; Diesendruck & Markson, 2001). An alternative proposal is that children have a constraint on the types of lexicons they consider when learning the meaning of a new word — namely, only those lexicons that have a one-to-one mapping between words and objects (Markman & Wachtel, 1988; Markman et al., 2003). One way to think about these different proposals is by the timescales over which they operate. A pragmatic constraint is a bias that relies on information available at the pragmatic or discourse timescale, while a one-to-one constraint is a bias that could be learned through experience over the developmental timescale. Using a hierarchical Bayesian model, we instantiate the pragmatic account through basic probabilistic properties of the model (this corresponds to they intuition: "If the novel word mapped to the familiar object, that would make it really unlikely I had heard a different label for that familiar object so many times before!"). We instantiate the lexical account by constraining the set of lexicons the learner considers. We show that, in principle, both sources of information, at two different timescales, pull in the same direction and lead the learner to select the novel object. This highlights an empirical challenge in trying to disentangle the relative contributions of information at each timescale to this inference.

A key component of learning is *generalization*: aggregating across tokens of examples observed over time in order to make predictions in new contexts. In acquiring language, children learn to generalize both the form and meaning of language. The are a number of forms these generalization could take. One possibility is that the form is simply a frequency count. For example, in the case of word learning, the child might track the number of times a word and object co-occur in the environment and then infer that the meaning of the word is the object that the word most often co-occurs with (what is often referred to as "cross-situational learning;"

Pinker, 1984; Yu & Smith, 2007; Smith & Yu, 2008).[8] This is slightly different than what is typically meant by "generalization," but it has the critical character of aggregating across tokens in order to make a prediction in a new context (e.g. that "dog" maps to DOG).

A second possible form of this generalization is an abstract *overhypothesis*. Overhypotheses have been called many things in the literature (e.g., theories, rules, schemas), but the key feature of this form of representation is that it is abstracted away from any particular observation. An example of such an overhypothesis is the understanding that shape is often the organizing feature of early word meanings (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002; Kemp, Perfors, & Tenenbaum, 2007). There is evidence that children as young as 9-months old can learn very simple overhypotheses (Dewar & Xu, 2010). In understanding the dynamics of this abstraction over time, hierarchical Bayesian models have provided a powerful framework for thinking about these abstractions. The key insight from these models is that the induction of an abstraction happens with only very few instances, and that high-level abstractions may be learned more quickly than lower-level abstractions (Goodman, Ullman, & Tenenbaum, 2011). This provides a promising suggestion for how children might learn the regularities of a language from impoverished input.

Regardless of the form that the generalization takes, the connection between the discourse and developmental timescales lies in the aggregation of instances of language use across time. In aggregating across instances, learners form some sort of generalization that allows them to make predictions in new contexts. This gives rise to two different cognitive processes: a discourse-based process in the experience of particular tokens of language use and a developmental-based process in the generalization of these tokens across time.

[8]Note that there is another generalization problem embedded in this problem: how to categorize objects as the same across contexts (M. Lewis & Frank, 2013a). For example, this is the problem of recognizing that a dalmatian and a terrier are both instances of the same category DOG. This problem must be solved jointly with the mapping problem.

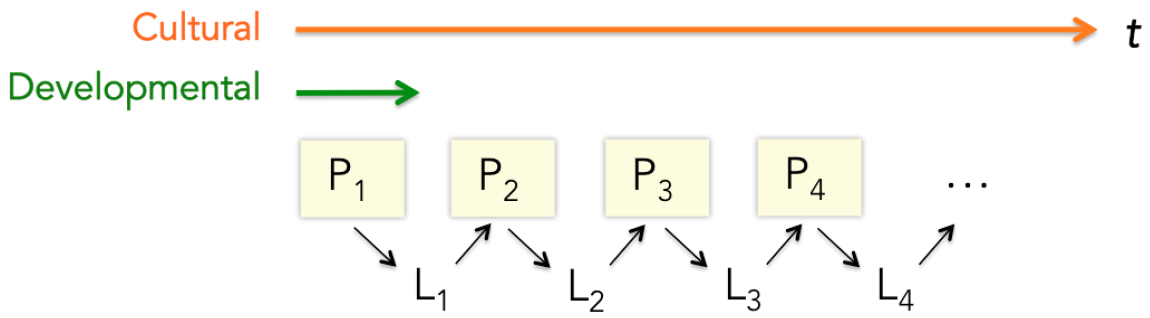

Figure 7. Relationship between the developmental and cultural timescales. The proposal is that each generation of speakers, P_n, develops structural regularities that define a language, L_n. This language then becomes the input for the next generation of speakers, $P_n + 1$. This iterative process leads to regularization in L over the cultural timescale.

Dynamics between developmental and cultural timescales

Generations of speakers acquiring and using language lead to the emergence of structure on the cultural timescale. These are the structural regularities discussed in Part I. This claim is the core of the argument presented by Christiansen and Chater (2008), which is summarized in Figure 7. The idea is that linguistic regularities emerge from groups of speakers playing linguistic coordination games over time. These regularities emerge in part because all speakers have the same pragmatic biases (as outlined by Horn (1984)). Each generation of speakers produces their own set of structural regularities (or, what can simply be referred to as "language"). Language change happens in the course of new speakers acquiring the language. This change is in the form of increasing regularity in the language. This more regularized language then becomes the input to the next generation of speakers. Thus, language, as a set of structural regularities, is never a static entity, but rather a constantly evolving entity over the cultural timescale.

There is evidence for this process of change in the laboratory. These experiments explore

the dynamics of language change through iterated learning experiments. Each participant in these experiments is trained and tested on an aspect of an artificial language. The key feature of these experiments is that the testing output of nth participant is used as the training input for the $(n+1)$th participant. Kirby, Cornish, and Smith (2008) provide an elegant demonstration of iterated learning in the laboratory. In their study, participants were presented with a novel language and asked to learn the pairings between words and novel images. For the first subject, the mappings between words and images were randomly generated. In the training phase, the subject was presented with an image and a label for that image. In the testing phase, they were shown a new image and asked to guess the word's meaning. This subject's responses were recorded, then divided into half (one half for training and one half for testing), and presented to a subsequent subject. The results revealed that as the number of generations increased, the language became more systematic in its mapping between features of the stimuli (e.g. color, shape, etc.) and syllables in the novel words. This provides evidence for the emergence of regularity over the cultural timescale.

The central issue in this theory is how and why regularity in structure emerges over the cultural timescale. The argument goes as follows. Language learners can only ever observe a subset of the language, and this creates a "bottleneck" in the transmission process (Kirby, 2007). Consequently, a language can only be acquired if it can be learned through impoverished data. Critically, the only way to learn a whole language from limited input is through generalization. Put another way, the only way that a language can be transmitted through a bottleneck is through compression into generalized rules. Thus, in the course of their acquisition, learners change the messy input they receive to become more regular. This has the consequence of making the language easier to learn for subsequent generations. Through simulations, Brighton, Smith, and Kirby (2005) demonstrate that the size of the bottleneck is directly related to the degree of

generalizations: when learners are given more input, they generalize less.

There is growing coherence in the empirical data around this view. There is evidence from several natural contexts that suggests that children are biased to regularize their linguistic input (e.g., Goldin-Meadow & Mylander, 1983; Senghas & Coppola, 2001). However, this result stands in contrast with work with adults. A number of studies suggest that in artificial learning experiments, adults reproduce unpredictability in their input, rather than regularize it (e.g., Hudson Kam & Newport, 2005). Why would we expect adults to behave differently than children? Recent computational work suggests a resolution to this inconsistency. Reali and Griffiths (2009) find that the regularity that emerges over generations ultimately reflects the distribution of prior hypotheses of individual language learners. Critically, how fast the language converges to this regularity over generations depends on the strength of the prior. Thus, one way to think about the empirical difference between children and adults in these language learning experiments is in terms of their distribution over priors: children have weaker priors. However, with enough transmission of the language across generations, all languages should converge on the prior. Indeed this is what they find in an iterated learning experiment. While individual adult learners match the variability in their input in a single generation, iterations of learning eventually leads to regularization.

Conclusion

Language is an incredibly complex phenomenon, as evidenced by the wide range of perspectives on how and what to study. In the present paper, I have tried to suggest a unifying theory for thinking about language use and language structure. I have suggested a causal story for how equilibria at the level of language use can lead to systematicity at the level of linguistic structure — and, in particular, systematicity that reflects the equilibria at the pragmatic timescale. By starting with Horn's pragmatic framework, I motivated a wide range of pragmatic phenomena.

The key claim is that local dynamics between adjacent timescales lead to the emergence of structure at the cultural timescale. While my focus has been the role of pragmatics, it is important to note that the claim is not a reduction of all linguistic phenomena to pragmatic principles. Rather, other factors like cognitive limitations are also likely to be important factors in the complete causal story.

The heart of the proposed theory is the micro-level processes that occur within the individual cognitive system. Understanding these processes, and how they lead to phenomena at the pragmatic, discourse and developmental timescales, is essential if we are to develop a complete understanding of language at every level. But, to know how these cognitive processes are related to broader linguistic phenomena, we must have a theory of how these phenomena are related to each other. That is the goal of the present paper. While it may be convenient and scientifically necessary to limit our focus of study to particular aspects of language, a broader theory of how these aspects are related to each other is important for guiding our enquiry.

However, testing the predictions of this theory is not an easy empirical task. The proposal is a long and complex causal chain of events leading to the emergence of linguistic structure. A large part of the empirical challenge arises from the fact that the target phenomena unfold across time, across varying timescales, and because the phenomena cannot be observed in any individual alone. Furthermore, in cases where we do have evidence for aspects of this causal chain, it is only for the dynamics between one or two timescales. For this reason, computational models will provide an invaluable tool in future work for making testable predictions from this complex causal theory.

References

Baddeley, R., & Attewell, D. (2009). The relationship between language and the environment: Information theory shows why we have only three lightness terms. *Psychological Science*, *20*, 1100–1107.

Baldwin, D. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62*, 874–890.

Bergen, L., Goodman, N. D., & Levy, R. (2012). Thats what she (could have) said: How alternative utterances affect language use. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society.*

Bergen, L., Levy, R., & Goodman, N. D. (in prep). Pragmatic reasoning through semantic inference.

Bion, R., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, *126*, 39-53.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*, B13–B25.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1482.

Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, *2*, 177–226.

Brueckner, J. K. (1987). The structure of urban equilibria: A unified treatment of the Muth-Mills model. *Handbook of Regional and Urban Economics*, *2*, 821–845.

Bybee, J., & McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*, *22*,

381–410.

Chomsky, N. (1965). Aspects of the theory of syntax cambridge. *Multilingual Matters: MIT Press*.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–509.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, *55*, 591–621.

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition. Hillsdale, NJ: Erlbaum*.

Clark, E. (1988). On the logic of contrast. *Journal of Child Language*, *15*, 317–335.

Clark, H. H. (1996). *Using language*. Cambridge University Press Cambridge.

Clark, H. H. (2006). Social actions, social commitments. *The roots of human sociality: Culture, cognition, and interaction*, 126–150.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, *22*, 245–258.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

Dewar, K., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge. *Psychological Science*, *21*, 1871.

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental Psychology*, *37*, 630.

Dryer, M. S. (2005). Order of subject, object, and verb. *The World Atlas of Language Structures*, 330–333.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their

input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*, 17897–17902.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*, 296–340.

Frank, M. C., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998–998.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 90-96.

Frawley, W. (2003). International encyclopedia of linguistics. In (Vol. 2, chap. Deixis.). Oxford University Press.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, *24*, 1079–1088.

Goldin-Meadow, S., & Mylander, C. (1983). Gestural communication in deaf children: noneffect of parental input on language development. *Science*, *221*, 372–374.

Goodman, N., Ullman, T., & Tenenbaum, J. (2011). Learning a theory of causality. *Psychological Review*, *118*, 110.

Grice, H. (1975). Logic and conversation. *1975*, 41–58.

Hockett, C. (1960). The origin of speech. *Scientific American*, *203*, 88-96.

Hopper, P. (1987). Emergent grammar. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* (Vol. 13, p. 139-157).

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, form, and use in context*, *42*.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and*

Development, *1*, 151–195.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*, 307-321.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Kiparsky, P. (1983). Word-formation and the lexicon. In *Proceedings of the 1982 Mid-America Linguistics Conference* (Vol. 3, p. 22).

Kirby, S. (2007). The evolution of language. *Oxford Handbook of Evolutionary Psychology*, 669–681.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*, 10681–10686.

Labov, W. (1972). 13. the social stratification of (R) in New York City department stores.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Lewis, D. (1969). *Convention: A philosophical study.* Harvard University Press, Cambridge, Mass.

Lewis, M., & Frank, M. C. (2013a). An integrated model of concept learning and word-concept mapping. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Lewis, M., & Frank, M. C. (2013b). Modeling disambiguation in word learning via multiple probabilistic constraints. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Lewis, M., Sugarman, E., & Frank, M. C. (2014). The structure of the lexicon reflects principles of communication. In *Proceedings of the 36th Annual Meeting of the Cognitive Science*

Society.

Mahowald, K., Fedorenko, E., Piantadosi, S., & Gibson, E. (2012). Info/information theory: speakers actively choose shorter words in predictable contexts. *Cognition*, *126*, 313–318.

Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Markman, E., Wasow, J., & Hansen, M. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*, 241–275.

McMurray, B., Horst, J., & Samuelson, L. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*, 831.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*, 201–213.

Mills, E. S. (1967). An aggregative model of resource allocation in a metropolitan area. *The American Economic Review*, 197–210.

Peirce, C. (1931). *The collected papers of Charles S. Peirce* (P. W. C. Hartshorne & A. Burks, Eds.). Cambridge: Harvard University Press.

Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*, 3526–3529.

Piantadosi, S., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280-291.

Pinker, S. (1984). *Language learnability and language development, with new commentary by the author* (Vol. 7). Harvard University Press.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*, 317–328.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*, 1436–1441.

Regier, T., Kemp, C., & Kay, P. (2014). Word meanings across languages support efficient communication.

Saussure, F. (1916). Course in general linguistics, trans. *London: Peter Owen*.

Schelling, T. C. (1980). *The strategy of conflict*. Harvard University Press, Cambridge, MA.

Senghas, A., & Coppola, M. (2001). Children creating language: How nicaraguan sign language acquired a spatial grammar. *Psychological Science*, *12*, 323–328.

Senghas, A., Kita, S., & Özyürek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, *305*, 1779–1782.

Sherif, M. (1935). A study of some social factors in perception: Chapter 3. *Archives of Psychology*, *27*, 23-46.

Smith, L., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*, 13-19.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, *27*, 979–1001.

Whorf, B. L. (1956). Language, thought, and reality: Selected writings of Benjamin Lee Whorf. *Cambridge, MA*.

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in bayesian word learning. *Developmental Science*, *10*, 288-297.

Xu, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Proceedings of the 35th Annual Meeting of the*

Cognitive Science Society.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.

Zipf, G. (1936). *The psychobiology of language*. Routledge, London.

Zipf, G. (1949). *Human behavior and the principle of least effort*. Cambridge, MA.