Application of Dynamical Systems to Model Human Language Development

Jeremy Irvin
[Advisor: Fermín Moscoso del Prado Martín]

Abstract

The theory of dynamical systems studies the interactions between multiple measurable variables that evolve over time [1]. Dynamical systems are described as systems of coupled differential equations that model this temporal coevolution of several variables. These techniques have been successfully applied in several areas of science such as ecology, economics, and meteorology. A particularly useful development is Takens' Delay Embedding Theorem [1, 2, 3, 4], which enables the approximate reconstruction of the state space using only a subset of the variables which are relevant to the system.

Multiple measurable linguistic variables have been found to be relevant for the study of human language acquisition and development [5]. Researchers are interested in understanding the relationships between such measures insomuch those are informative of the mental processes and structures involved. A recurrent problem in this field is that not all the relevant variables are available to the researcher. Even more problematic is the fact that not only are these values unknown, but the identities of the variables themselves are unknown. Takens' Theorem opens the possibility of detailing the structure of the system, eschewing the need to explicitly identify all relevant variables.

In this thesis, I intend to apply these dynamical system techniques to language acquisition data [6]. Using linguistic corpus data from the CHILDES database ([7]; a large collection of transcribed conversations between children of different ages and their caretakers), I intend to model the interactions between indicators of different aspects of language acquisition. These indicators include measures such as vocabulary richness, diversity of syntactic structures, and degree of word internal structures. These variables are known to exhibit complex nonlinear patterns during a child's development. The goal of this thesis is to investigate whether such nonlinear patterns can arise from a relatively simple system of differential equations, and to use recent causal modeling techniques to infer which variables exhibit meaningful interactions [8]. Were this the case, this would enable us to infer additional properties of the system such as capacity limitations, etc. [6]. Finally, it is quite possible that the same factors that give rise to the nonlinear patterns during acquisition (infancy) might themselves help make predictions on the deterioration of the language system during old age [5].

Contents

List of Figures

List of Tables

1 Introduction

There have been previous efforts to understand the complex interaction within the child-mother dyad. This and other relevant linguistic background is provided first in the next section. Next background on chaotic time series, dynamical systems, differential topology, and Taken's Theorem(s) will be provided in order to understand the methods behind the analysis. Then the multiple comparisons problem and the false discovery rate will be explained, in addition to a brief tangent into a basic clustering algorithm used later in the analysis. After this background, an outline of the analysis is described in detail, followed by a discussion of the findings. Finally potential future directions will be proposed.

2 Background, Preliminary, and Related Work

2.1 Child Language Acquisition

Child-directed language (CDL) –sometimes referred to as "motherese"– is the pattern of language used by parents when talking to young children. It is known to exhibit distinctive characteristics with respect to regular adult language (cf., [11]): It typically uses shorter utterances, prosody is often exaggerated, redundancy and repetition are higher than normal, and referential context tends to be linked to very immediate contexts. It has been observed that the lexical and syntactic complexity of CDL gradually increases along a child's development [12, 13], eventually converging to regular adult language. Whether the simplicity of CDL relative to regular adult language plays a functional role in facilitating language acquisition is a contentious issue in the literature. Some researchers argue that "starting small" [14] is a fundamental aspect that facilitates language acquisition (e.g., [15]). Others, however, claim that there is no facilitating role played by such simplicity (e.g., [16]). This latter group would claim that, to some degree, children might exploit universal aspects of language structure, and their syntactic performance would be unrelated to the input to which they have been exposed. Those researchers advocating for a functional role of the simplified input refer to *fine-tuning* [13, 17, 18] as the process by which caregivers adjust the complexity of CDL as a function of the level of complexity of the language produced or understood by the child.

A weak and a strong version of the fine-tuning hypothesis compete in the literature. In the weak interpretation, although parents gradually increase the complexity of their CDL, they do not do so as a direct response to the specific properties of the utterances produced by their children, but rather they adjust to the children's overall level of cognitive development, irrespective of the specificities of the language they produce and understand. In this line, several studies have failed to find a direct link between the complexity of the parent's language and that of the child's [19, 20, 21]. These findings suggest that –if anything– the complexity of CDL might increase as a function of the child's age or overall level of development, but not so much as a direct response to the detailed properties of CL. In contrast, other researchers have found evidence supporting a strong version of the fine-tuning hypothesis: That parents adapt the complexity of CDL in direct response to the specific properties of CL [17, 18, 22, 23, 24, 25]. The strong version of the fine-tuning hypothesis has become the dominant view in the field, considered a well-established fact by influential researchers (e.g., [26]).

2.2 Linguistic Measures and Entropy

2.2.1 Shannon's Entropy and Entropy Bias

Several useful linguistic measures with foundations in information theory have been proposed that are relevant to the analysis of linguistic development [5]. Essential to each of the measures is the concept of entropy, originally proposed by Shannon [27]. Generally, given a script of characters X, Shannon's entropy is defined as

$$H[X] = -\sum_{x \in X} p(x) \log p(x).$$

where $p(x)$ denotes the probability of a certain character x in the script. This value represents a degree of uncertainty. In most ecological (and linguistic) applications, a higher degree of uncertainty implies

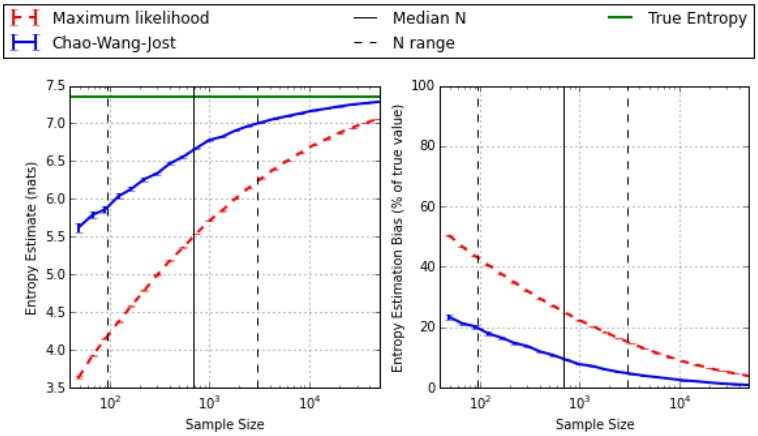

Figure 1: [5] Average estimator values and biases of the entropy estimators using maximum likelihood (Shannon's equation, red lines) and CWJ method (blue lines) relative to the true entropy (green line) as a function of the sample size. The left-hand panel plots the actual estimator values and the right-hand panel plots the biases relative to the true entropy. The horizontal axes use a logarithmic scale.

more diversity, so this value represents the diversity of characters in the script.

In practice, the total script size $|X|$ is unknown, so this entropy must be approximated. However, estimating this quantity leads to significant underestimates of the true entropy when small amounts of data are available. This computed value is a consistent estimator of the true entropy, meaning it will converge with probability 1 to the true value as the sample size increases. However, it is biased - although it will eventually reach this true value, the estimator will significantly underestimate this value. This is known as the problem of entropy bias, and was first discovered by Miller [28]. Fig 1 illustrates this problem arising in data from the Switchboard Corpus, a collection of telephone conversation transcriptions. The true entropy in this example is computable because the data being estimated was generated from a known (multinomial) distribution for each sample size.

In an attempt to alleviate this entropy bias problem, numerous methods have been proposed that produce a less biased entropy estimate than the observed sample entropy. These methods include the traditional bias-reduction method, the jackknife estimator [29], coverage-adjusted method [30], the Grassberger estimator [31], and many more. The most recent and successful estimator is the reduced bias entropy estimator is proposed by Chao, Wang, and Jost [30]. Their work links species accumulation curve slopes and entropy, allowing for an alternate (less biased) method of estimating the entropy using these slopes. An in detail derivation of this technique can be found in [30]. The CWJ method seems to produce a good (low bias) entropy estimator compared to existing methods given the total population size is unknown. However, this estimate is still significantly biased in small sample sizes, as seen in Fig. 1.

2.2.2 Lexical Diversity

The first linguistic measure, known as *lexical diversity*, denoted $H[L]$ is the entropy of the frequency distribution of lemmas,

$$H[L] = -\sum_{\ell \in L} p(\ell) \log p(\ell),$$

where L is the set of all lemmas and $p(\ell)$ denotes the probability of the speaker using a particular lemma $\ell \in L$. This measure represents the diversity of the vocabulary (lexicon) of the speaker, hence the name lexical diversity. The set of all possible lemmas available for use by an individual speaker is unknown, so this value must be approximated using the CWJ technique.

2.2.3 Inflectional Diversity and Syntactic Diversity

Similarly, the (corrected) entropy of the frequency distribution of unlemmatized words, denoted $H[W, L]$, is used to compute the *inflectional diversity*, denoted $H[W|L]$, which is the difference between these two entropies:

$$H[W|L] = H[W, L] - H[L].$$

This measure attempts to capture the complexity of the speaker's inflections, or word form changes. The final measure that will be used is known as the *syntactic diversity*, a measure of the speaker's grammatical complexity. This can be calculated by inducing a probabilistic context-free grammar (PCFG) from data-linked syntactic parse trees by maximum likelihood estimation and computing the entropy of the parse trees it generates. However, this measure is highly correlated with the speaker's *mean length of utterance* (MLU) [5], which is much easier to compute. Thus MLU acts as a simple and valid replacement for determining the syntactic diversity of the speaker.

2.3 Dynamical Systems Models and Causality Detection

Sequence of data points consisting of consecutive measurements over time are known as time series. These measurements are often irregular - plagued with noise and highly nonlinear. Dynamical systems are sets of relationships between multiple measurable quantities which can be expressed by a set of rules, most often a system of differential equations. The goal of studying chaotic time series is to extract meaning amongst the noise and create a model in order to make future predictions on the data. Linear systems are of course much more well-understood than nonlinear systems, but rarely do linear systems arise in practice.

2.3.1 Cognitive Science for Modeling Language Growth

Dynamical systems[1] offer powerful tools for modeling human development (e.g.,[32, 6, 33]). These models provide a mathematical framework for implementing the principle that development involves the mutual and continuous interaction of multiple levels of the developing system, which simultaneously unfold over many time-scales. Typically, a dynamical system is described by a system of coupled differential equations governing the temporal evolution of multiple parts of the system and their interrelations. In recent years, it has been noticed that, in human development, such systems extend beyond the individual. In particular, it has been found that the linguistic and behavioral interaction between parent-child dyads can be jointly considered as part of a single dynamical system encompassing both the child and the parent [34, 35]. In this direction [6] shows that the interaction of components within the child-parent dyad can also be modeled on the time-scale of development itself. One difficulty that arises when trying to model a dynamical system as complex as the joint development of language in a parent–child dyad is that many factors that are important for the evolution of the system might not be available or might not be easily measurable or –even worse– there are additional variables relevant for the system of which the modeler is not even aware. In this respect, a crucial development was the discovery that, in a deterministic coupled dynamical system –even in the presence of noise– the dynamics of the whole system can be satisfactorily recovered using measurements of a single of the system's variables (Takens' Embedding Theorem; [3]).

The finding above opens an interesting avenue for understanding the processes involved in language acquisition (perhaps more suitably termed language *growth*, following [6]). In the same way that systems of differential equations can be used to model the evolution of ecosystems (e.g., predator-prey systems), one could take measurements of the detailed properties of CL and CDL, and build a detailed system of equations capturing the macroscopic dynamics of the process. However, in order to achieve this, it is necessary to ascertain the ways in which different measured variables in the system affect each other. This problem goes beyond estimating correlations (as could be obtained, for instance, using regression models), as one needs to detect asymmetrical *causal* relations between the variables of interest, so that these causal influences can be incorporated into the models.

[1] There is some inconsistency in the literature in the use of the terms "dynamic systems" and "dynamical systems". Here, we follow [32] in using the latter to refer to the specific implementation of the former using systems of coupled differential equations.

2.3.2 Taken's Theorem(s)

There are several tools that are typically used in the analysis of nonlinear dynamical systems, the most notable being Taken's result [1, 3]. The space in which the orbits of a dynamical system evolved is known as state space. Given a sequence of scalar measurements from a single variable, we wish to reconstruct the state space of the system which produced these measurements. If we could discover the values of time derivatives of the measured variable, we could find a connection between the derivatives and the state variables, ie, the differential equations which produced the observations. But this is very difficult (naive estimations of the derivatives tend to be very inaccurate). Moreover, variables relevant to the underlying state space are often either unknown or infinite in number, so discovering the underlying attractor which describes these variables seems even more problematic.

The breakthrough that allows one to extract information about state space from time series on an evolving system is the following: to capture the orbits in state space, we directly use time lagged variables of the initial measurements. Then, for every candidate time index n, we use a collection of d composed time lags T to create a d-dimensional vector $y(n)$ that represents the underlying manifold in state space which produced the original time series [1]. If $s : M \to \mathbb{R}$ is a time series defined on the set of time indices M, then the vector constructed is

$$y(n) = \left[s(n),\, s(n+T),\, s(n+2T),\, \cdots,\, s(n+(d-1)T)\right].$$

Taken's result [2, 3], known as the *method of delays*, provides a solid theoretical foundation to this intuition, namely that any smooth nonlinear change of variables will suffice to produce a manifold which retains many of the invariant properties of the true underlying (perhaps infinite dimensional) manifold generating the observations. The following is a brief tangent into basic differential topology [4] in order to formulate the statement of Takens' Theorem.

Definition. A *topological space* is a pair $(X, \mathcal{T})$ where X is a set and $\mathcal{T}$ is a collection of subsets of X with

- $\emptyset, X \in \mathcal{T}$,
- $T_1 \cap T_2 \in \mathcal{T}$, $\forall T_1, T_2 \in \mathcal{T}$, and
- $\bigcup_{\alpha \in \Omega} T_\alpha \in \mathcal{T}$, for any (potentially uncountable) index set Ω.

We say that $\mathcal{T}$ is a topology on X, and elements of $\mathcal{T}$ are called *open sets*.

Example. Consider $X = \mathbb{R}^n$ with standard Euclidean metric d defined by $d(x, y) = ||x - y||$ where $||\cdot||$ defines the standard Euclidean norm. Define the ball $N_\epsilon(x)$ of radius ϵ around x to be

$$N_\epsilon(x) = \{y \in X : d(x, y) < \epsilon\}.$$

Say that $U \subseteq X$ is in $\mathcal{T}$ if $\exists \epsilon > 0$ s.t. $N_\epsilon(x) \subseteq U$ for all $x \in U$.
It is easily verified that $(X, \mathcal{T})$ satisfies the above axioms of a topological space. Moreover, $N_\epsilon(x)$ is open for every $x \in \mathbb{R}^n$, $\epsilon > 0$, so $N_\epsilon(x)$ is called an open ball around x of radius ϵ. This $\mathcal{T}$ is called the *standard topology* on $\mathbb{R}^n$.

Definition. Let $(X_1, \mathcal{T}_1)$ and $(X_2, \mathcal{T}_2)$ be two topological spaces. A function $f : X_1 \to X_2$ is *continuous* if $f^{-1}(U_2) \in \mathcal{T}_1$ for all $U_2 \in \mathcal{T}_2$.

Definition. Let $(X_1, \mathcal{T}_1)$ and $(X_2, \mathcal{T}_2)$ be two topological spaces. A function $f : X_1 \to X_2$ is *homeomorphic* if f is continuous, bijective, and f^{-1} is continuous.

Definition. Let $E \subseteq \mathbb{R}^n$ be open (with the topology induced from the standard topology of $\mathbb{R}^n$), and $f : E \to \mathbb{R}^m$. Then f is *differentiable* at $x \in E$ if there exists $Df(x) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ s.t.

$$\lim_{h \to 0} \frac{||f(x+h) - f(x) - Df(x)h||_{\mathbb{R}^m}}{||h||_{\mathbb{R}^n}} = 0.$$

$Df(x)$ is called the *derivative* of f at x. Since this mapping is linear, it can be represented as a matrix, known as the *Jacobi matrix* of f at x, which is a matrix containing the partial derivatives of f.

Definition. Let $(X_1, \mathcal{T}_1)$ and $(X_2, \mathcal{T}_2)$ be two topological spaces. A function $f : X_1 \to X_2$ is *diffeomorphic* if f is differentiable, bijective, and f^{-1} is differentiable.

Definition. Let $E \subseteq \mathbb{R}^n$ be open. The set of all n times differentiable functions on E whose nth derivative is continuous is called $\mathcal{C}^n(E)$. The set of all infinitely differentiable functions is called $\mathcal{C}^\infty(E)$. If $f \in \mathcal{C}^\infty(E)$, it is said to be *smooth*.

Definition. A topological space $(X, \mathcal{T})$ is *Hausdorff* if for all distinct $p, q \in X$, there exists $U, V \in \mathcal{T}$ such that $p \in U$, $q \in V$, and $U \cap V = \emptyset$.

Definition. Let $(X, \mathcal{T})$ be a topological space. A subset $\mathcal{B} \subseteq \mathcal{T}$ is called a *countable base* for $\mathcal{T}$ if $\mathcal{B}$ is countable and every element of of $\mathcal{T}$ is the (not necessarily finite) union of elements in $\mathcal{B}$.

Definition. If a topological space $(X, \mathcal{T})$ has a countable base, then it is called *second countable*.

Definition. An *n-dimensional topological manifold* is a Hausdorff, second countable topological space $(M, \mathcal{T})$ such that every $p \in M$ lies in an open set which is homeomorphic to an open set of $\mathbb{R}^n$, with the topology induced from the standard topology of $\mathbb{R}^n$. We will often drop the $\mathcal{T}$ and simply call M an n-dimensional topological manifold (or a topological manifold of dimension n).

Definition. Let M be an n-dimensional topological manifold. A *chart* (or *coordinate system*) on M is a pair (U, ϕ) where U is an open subset of M and ϕ is a homeomorphism from U onto an open subset $\phi(U)$ of $\mathbb{R}^n$ with U the *chart domain* (or *coordinate neighborhood*) and ϕ the *coordinate function*.

Definition. An *atlas* $\mathcal{A}$ on M is a collection $\mathcal{A} = \{U_\alpha, \phi_\alpha) : \alpha \in \Omega\}$ of charts on M such that

$$\bigcup_{\alpha \in \Omega} U_\alpha = M.$$

Definition. Let (U, ϕ) and (V, ψ) be two charts of an n-dimensional manifold M with overlapping domains. Then the functions from open subsets of $\mathbb{R}^n$ to $\mathbb{R}^n$

$$\phi \circ \psi^{-1} : \psi(U \cap V) \to \mathbb{R}^n \text{ and } \psi \circ \phi^{-1} : \psi(U \cap V) \to \mathbb{R}^n$$

are called *coordinate transformations*.

Definition. Two charts are *$\mathcal{C}^r$-related* if their coordinate transformations are $\mathcal{C}^r$.

Definition. An atlas is *$\mathcal{C}^r$-differentiable* its charts are pairwise $\mathcal{C}^r$-related.

Definition. A *differential structure* is the set of all charts which are $\mathcal{C}^r$-related to those in a particular atlas.

Theorem. A differential structure is an atlas.

Definition. A manifold with a differential structure is *differentiable*, and is called a *$\mathcal{C}^r$ manifold*.

Definition. Let M and N be two n-dimensional manifolds. A function $f : M \to N$ between $\mathcal{C}^r$ manifolds is *$\mathcal{C}^s$-differentiable* $(s \leq r)$ if for every $p \in M$, there exist charts (U, ϕ) and (V, ψ) of M and N respectively with $p \in U$ and $f(p) \in V$ such that $\phi \circ f \circ \psi^{-1} : \psi(U \cap f^{-1}(V)) \to \mathbb{R}^n$ is $\mathcal{C}^s$.

Remark. The Jacobi matrix of f at p with respect to ϕ and ψ is $D\phi f \psi^{-1}(\psi(p))$. This matrix depends on the charts used, but its rank does not.

Definition. Let M and N be two $\mathcal{C}^r$ n-dimensional manifolds, and $f : M \to N$ be $\mathcal{C}^s$-differentiable. If the $D\phi f\psi^{-1}(\psi(p))$ is injective, then f is *immersive* at p, if f is immersive at every $p \in M$ is an *immersion.* A homeomorphic immersion which maps M onto its image is an *embedding.* If $D\phi f\psi^{-1}(\psi(p))$ is surjective, then f is *submersive* at p.

[2]**Theorem 1 (Takens).** Let M be a compact manifold of dimension m. For pairs (ϕ, y) with $\phi \in Diff^2(M)$, $y \in \mathcal{C}^2(M, \mathbb{R})$, it is a generic property that the map $\Phi_{(\phi,y)} : M \to \mathbb{R}^{2m+1}$, defined by

$$\Phi_{(\phi,y)}(x) = (y(x), y(\phi(x)), ..., y(\phi^{2m}(x)))$$

is an embedding.

Here, 'generic' means open and dense relative to the $\mathcal{C}^1$ topology.

[2]**Theorem 2 (Takens).** Let M be a compact manifold of dimension m, and $\phi : M \to M$ a diffeomorphism with the properties that

- there are a finite number of periodic points of ϕ with periods less than or equal to $2m$, and
- if x is any periodic point with period $k \leq 2m$, then the eigenvalue of the derivative of ϕ^k are all distinct.

Then for generic $y \in \mathcal{C}^2(M, \mathbb{R})$, the map $\Phi_{(\phi,y)} : M \to \mathbb{R}^{2m+1}$ defined by

$$\Phi_{(\phi,y)}(x) = (y(x), y(\phi(x)), ..., y(\phi^{2m}(x)))$$

is an embedding.

See [2] for an in detail walkthrough of the proofs of these theorems.

2.3.3 Intuition Behind Taken's Theorem

Taken's theorems rely on some relatively heavy differential topology, but they have much simpler implications on the study of chaotic time series. We assume the state at any time t_{i+1} is uniquely determined by the state at the previous time t_i (namely the Markov property), and also that the reverse is true (this change is invertible). This provides the intuition for the invertible mapping $\phi : M \to M$ which maps the state at t_i to the state at t_{i+1}, with both ϕ and its inverse differentiable (ϕ differentiable). As stated before, often the dynamics stem from a system of differential equations, and in fact the condition that ϕ is diffeomorphic stems from differential equations theory. Typically $\phi(x) = x + T$ for some $T \in \mathbb{R}$ is used, which is clearly diffeomorphic. The measurements (the real valued outputs of the time series) are specified by the (twice) diffferentiable function $y : M \to \mathbb{R}$. We simply require that the measurement at the time t_i is uniquely specified by the state at time t_i, and this dependence is smooth ($\mathcal{C}^2$). Then the theorems say that if we use the time series and create vectors of $n > 2m$ elements (where m is the dimension of the underlying manifold M) with ϕ and y generic, then these vectors lie on an embedding of M in $\mathbb{R}^n$.

Constructing this embedding allows us to infer invariants of the system, including the dimension and topology of the state space, Lyapunov exponents, etc. Furthermore, the dynamics often do not explore all of its state space and are confined to subsets of the space called attractors. The reconstruction contains important information about these attractors as well. Calling $N = \Phi_{(\phi,y)}(M)$ the image of he state space under the embedding $\Phi_{(\phi,y)}$, then N is both a copy of the space M and the dynamics. The map $\psi : N \to N$ by $\psi = \Phi \circ \phi \circ \Phi^{-1}$ is a dynamical system (with domain N) which is also invertible and smooth. Notice that the periodic orbits of ϕ are mapped into corresponding periodic orbits of ψ by the embedding Φ. This is one of many properties preserved by the embedding.

2.3.4 Applying Taken's Theorem

Actual application of this topological machinery to real data is a very different task [1]. The first issue is the choice of the time delay T to be used in the embedding. The work of Taken's shows that any such

T will result in an appropriate embedding, but this is not the case in practice. If T is too small, then the adjacent coordinates will be so close numerically that they will be indistinguishable, and thus have not provided two independent coordinates. If T is too large, then the coordinates will be statistically independent, and the resulting embedding will be a projection onto two completely unrelated directions. So the choice of T will effect the resulting embedding. Multiple heuristics for the choice of T have been developed and tested that seem to work well in practice. The most successful prescription is to use the first minimum of the average mutual information between the two coordinates. Suppose $s : M \to \mathbb{R}$ is a time series. The average mutual information is the average amount of information we have about one measurement $s(n+T)$ given an observation $s(n)$, and vice versa, given by

$$I(T) = \sum_{n=1}^{N} P(s(n), s(n+T)) \log_2 \left[\frac{P(s(n), s(n+T))}{P(s(n))P(s(n+T))} \right].$$

Note that if $s(n)$ and $s(n+T)$ are statistically independent, then $I(T) = 0$, since this would imply $P(s(n), s(n+T)) = P(s(n))P(s(n+T))$. The first minimum of I yields coordinates which are independent but not probabilistically independent. If I has no minimum, T is usually taken to be 1 or 2. There is no general rule to choosing T that will yield the best result in every case - this is simply a choice that tends to work well in practice. Of course different choices of T should be tested when attempting to reconstruct the state space. In the analysis in this thesis, the average mutual information I was computed using the `tseriesChaos` R package [36].

Now that the time lag T has been chosen, the next step is to choose the embedding dimension d, ie, the number of coordinates to use in the embedding [1]. The intuition behind choosing this dimension is to find a Euclidean space $\mathbb{R}^d$ large enough so that the points on the attractor with dimension d_A can be unfolded (to create the embedding) without ambiguity. Two points which are near eachother in dimension d should be close because that is the property of those points, not because the dimension d is small. We can also think of the system as a group of orbits in high dimensional space (perhaps infinite), and the attractor as a small subset of the space with finite dimension d_A. We want to find a projection of the whole state space to a smaller subspace which captures the properties of the attractor. If the projection is too small, then observing evolution of the system in this space will be difficult due to incorrect neighboring points. Taken's theorem provides us with a sufficient dimension d_E for the unfolding of the attractor in the embedding space. However, this dimension is not always necessary, and often using the larger dimensions will result in much more costly computation due to the curse of dimensionality (in fact the computation cost rises exponentially with the dimension). There are several methods for choosing this embedding dimension, including singular value decomposition of the sample covariance matrix and false nearest neighbors. I will not go into detail in any of these methods, if more detail is desired please consult [1]. The embedding dimension is determined in this analysis using the `multispatial` R package [37], whose several other functionalities will be further discussed later.

2.3.5 Causality and Convergent Cross Mapping

By finding the time lag T and embedding dimension d, we have effectively reconstructed the state space from the original time series. We now wish to determine causality between the measured variables. Despite the commonly-held maxim that "correlation does not imply causation", recent work has found quite the opposite. [38] argues that, whenever two variables are correlated, there *must* exist some causal link between them. Namely, if variables A and B are found to be correlated, then one of three possibilities must be true: (a) A causes B, (b) B causes A, or (c) there is a third variable C causing both A and B. From this starting point, Pearl extended the probability calculus to account for causal relations. This approach offers a set of tools relying on Bayesian networks that enable –under some circumstances– to make inferences on the causal relations between measured variables. Relatedly, in economics, [39] introduced a technique that can recover asymmetric causal relations between pairs of time-series, what is usually termed Granger-causality or G-causality. Granger's techniques have an advantage over Pearl's in that they produce more reliable estimates when dealing with temporal data (e.g., [40]). As useful as these two approaches are, however, their power is severely limited by the fact that, in order to ascertain a *direct* causal relation between any two variables, one needs to discard

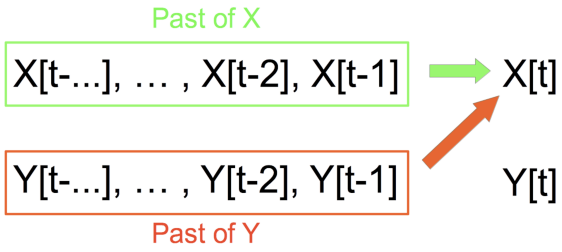

Figure 2: Granger Causality: variable Y Granger-causes variable X because the past of Y can provide a better prediction of future of X than solely using the past of X.

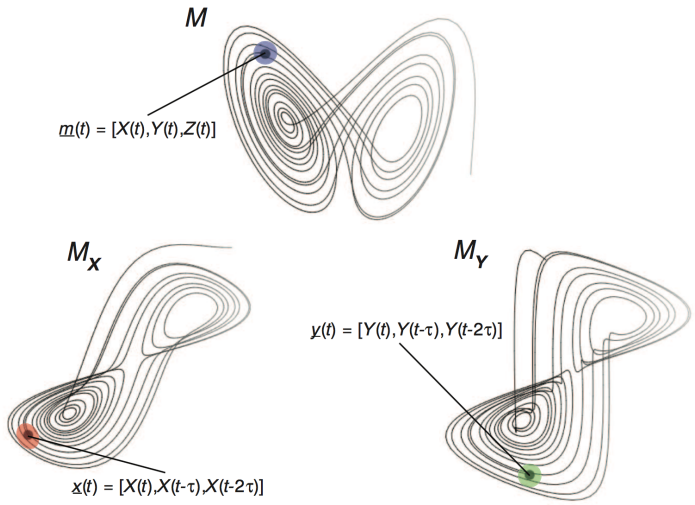

Figure 3: Reconstructed manifold for Lorenz's system (M; top), as well as the shadow manifolds reconstructed considering only X (M_X; bottom-left) and Y (M_Y; bottom-right) (reprinted with permission from [8]).

possibility (c) above, for which one requires knowledge of all other variables that might be of importance for the system in question. Without knowledge of these possible intervening variables, causal inference using these methods must remain strongly suspect. This problem is, of course, particularly acute when one is dealing with systems as complex as is human language.

Granger-causality relies on the notion of *separability*, this is, that the information contained by a causal source is unique to it, so that just eliminating that variable from consideration suffices for eliminating the information that it contributes. Purely stochastic systems often exhibit separability. Unfortunately, however, separability is not a property that is exhibited by deterministic non-linear dynamical systems. For studying the interactions of species within ecosystems, [8] introduced *Convergent Cross Mapping* (CCM), a causality-detection technique that is valid for non-separable systems, is capable of identifying weakly coupled variables even in the presence of noise, and –crucially– can distinguish direct causal relations between variables from effects of shared driving variables (i.e.,in possibility (c) from the previous paragraph, CCM would *not* find causality).

For instance, consider E. Lorenz's often studied dynamical system including three coupled variables

$X(t)$, $Y(t)$, and $Z(t)$ whose co-evolution is described by the system of differential equations

$$\begin{cases} \dfrac{\mathrm{d}X}{\mathrm{d}t} = \sigma(Y - X) \\ \dfrac{\mathrm{d}Y}{\mathrm{d}t} = X(\rho - Z) - Y \\ \dfrac{\mathrm{d}Z}{\mathrm{d}t} = XY - \beta Z \end{cases} . \tag{1}$$

The first equation in this system indicates that there is a relation by which Y causes X, as the change in X (i.e., its future value) depends on the value of Y (i.e., the future of X depends on the past of Y even after the past of X itself has been considered), a causal relation whose strength is indexed by parameter σ. The manifold defined by these three variables (Lorenz's famous strange attractor), which we can denote by M, is plotted in the top of Fig. 3. In many circumstances, however, not all variables of the system are available (some might be difficult to measure, or we might not even be aware of their relevance). It is at this point that Taken's Embedding Theorem comes into play. Informally speaking, the theorem states that the properties of a coupled dynamical system's attractor can be recovered using only measurements from a single one of its variables. This is achieved by considering multiple versions of the same variable lagged in time, that is, instead of plotting $(X[t], Y[t], Z[t])$, when only measurements of X are available, we can plot $(X[t], X[t+\tau], \ldots, X[t+(E-1)\tau])$. These reconstructed manifolds are termed "shadow" manifolds. M_X denotes the shadow manifold of M reconstructed on the basis of X alone. There are well-studied techniques for finding the appropriate values for the parameters for the lag τ and the number of dimensions E (c.f.[1]) so that the properties of the original manifold M are recovered by the shadow manifold M_X. Fig. 3 illustrates this point by plotting the shadow manifolds M_X (bottom-left) and M_Y (bottom-right) for the Lorenz system. Notice how both shadow manifolds recover much of the original's structure, using only knowledge of one of its three variables.

Each point in the original manifold M maps onto points in its shadow manifolds, as is illustrated by the points labeled $m(t)$, $x(t)$, and $y(t)$ in Fig. 3. The preservation of the topological properties of the original manifold in its shadow manifolds entails that points that are close-by in the original manifold will also be close-by in its shadow versions. This implies that, for causally linked variables within the same dynamical system, the state of one variable can identify the states of the others. [8] noticed that, when one variable X stochastically drives another variable Y, information about the states of X can be recovered from Y, but not vice-versa. This is the basic insight of the CCM method. To test for causality from X to Y, CCM looks for the signature of X in Y's time series by seeing whether the time indices of nearby points on M_Y can be used to identify nearby points on M_X. Crucially, in order to distinguish causation from mere correlation, CCM requires *convergence*, that is, that cross-mapped estimates improve in estimation accuracy with the sample size (i.e., "library size") used for reconstructing the manifolds. As the library size increases, the trajectories defining the manifolds fill in, resulting in closer nearest neighbors and declining estimation error, which is reflected in a higher correlation coefficient between the points in the neighborhoods of the shadow manifolds. Convergence then becomes the necessary condition for inferring causation. Using both artificial systems and ecological time-series with known dynamics, Sugihara and his colleagues demonstrated that this technique successfully recovers true directional causal relations when these are present, and –crucially– is able to discard spurious causation in the case when both variables are causally driven by a third, unknown, variable, but there is no true direct causation between them.

An inconvenience of CCM, and in general of techniques that rely on manifold reconstruction, is that they generally require that relatively long time-series of the behavior of the system are available. Such long series are, however, very difficult, if not impossible, to obtain in many fields, including of course language acquisition. One can however obtain multiple short time series from different instances of a similar dynamical system. In ecology, for instance, one can obtain short sequences of measurements of the population densities of a group of species measured at different places and times. In language acquisition, we might have multiple, relatively short longitudinal sequences of measurements from different children. With this in mind [9] developed Multispatial CCM (mCCM), an extension of CCM able to infer causal relations from multiple short time-series measured at different sites, making use

	Declared Non-significant	Declared Significant	Total
True null hypotheses	T_P (true positives)	F_P (false positives)	m
Non-true null hypotheses	F_N (false negatives)	T_N (true negatives)	$m - m_0$
	$m - R$	R	m

Table 1: [10] Relevant variables when testing m null hypotheses.

of dewdrop regression [41] to take the additional heterogeneity into account (these techniques are all implemented in the `multispatial` R package [37]). This procedure is outlined in detail in the appendix and supplementary material found in [9].

2.4 The Multiple Comparisons Problem and the False Discovery Rate

This thesis will provide an analysis of the causal relations between many different variables simultaneously. Determining causality requires hypothesis testing testing to ensure the causal links are statistically significant. We have no theory supporting the type of causal relations that we should expect to find, so we are essentially 'fishing' for relations. Performing many different statistical inferences at once without theoretical foundation can be very dangerous - when one evaluates a set of statistical inferences as a whole, it is much more likely for hypothesis tests to yield a Type I error (incorrectly rejecting the null hypothesis, meaning the tests will incorrectly determine there is a causal relation between the variables). This problem is known as the *multiple comparisons problem*. Formally, suppose we are testing m null hypotheses simultaneously, that m_0 of them are true, and that R is the number of hypotheses rejected. Table 1 summarizes the different relevant variables and their meanings.

R is an observable random variable, and T_P, F_P, F_N, and T_N are all unobservable random variables. The *per comparison error rate* (PCER) is defined as the expected proportion of false positives out of all hypotheses tested, $\mathbb{E}[F_P/m]$. The *familywise error rate* (FWER) is defined as the probability of at least one false positive, $P(F_P \geq 1)$. If we test each hypotheses with significance level α, then the FWER is at most α ($\mathbb{E}[F_P/m] \leq \alpha$). Similarly, if we test each hypotheses with significance level α/m, we are guaranteed that $P(V \geq 1) \leq \alpha$. The proportion of false positives to total errors R is a random variable $X = \frac{F_P}{R+F_P}$. The *false discovery rate* (FDR) is defined as the the expected proportion of falsely rejected hypotheses, $\mathbb{E}[X]$. The important thing to notice is that when $m_0 < m$ (at least one null-hypothesis is non-true), if the FWER is bounded (or controlled), then the FDR is bounded as well. However, a bound on the FDR does not imply a bound on the FWER, so FDR controlling procedures may be less strict then FWER controlling procedures, allowing for improved statistical inference.

There are many techniques which attempt to solve the issue of multiple comparisons, including the Bonferroni correction and *false discovery rate* (FDR) controlling procedures [10]. The Bonferroni correction tries to control the FWER directly, so it tends to be too strict (increasing the probability of false negatives). FDR controlling procedures, unlike the Bonferroni correction, are less conservative (they provide no direct bound on FWER), but they result in more Type 1 errors. However, this sacrifice tends to be insignificant and worth the resulting powerful statistical inference. The FDR controlling procedure proposed by Benjamini and Hochberg is an effective method of dealing with the the multiple comparisons problem even when the test statistics positively depend on the test statistics corresponding to the true null hypotheses [42].[2]

2.5 Clustering

Clustering methods attempt to discover groups of similar data given a series of data points. There are two general types of cluster analysis techniques, namely partitioning and hierarchical. Partitioning methods divide the dataset into k clusters where k is specified by the user. This is typically done by running the algorithm on a range of k values and choosing the k which yields the optimal clustering quality value, which is output by the algorithm. Hierarchical clustering discovers a hierarchy

[2]These methods are described in detail with examples in [10, 42] and are implemented in the `stats` package in R. They will be employed when searching for causal relations between time series.

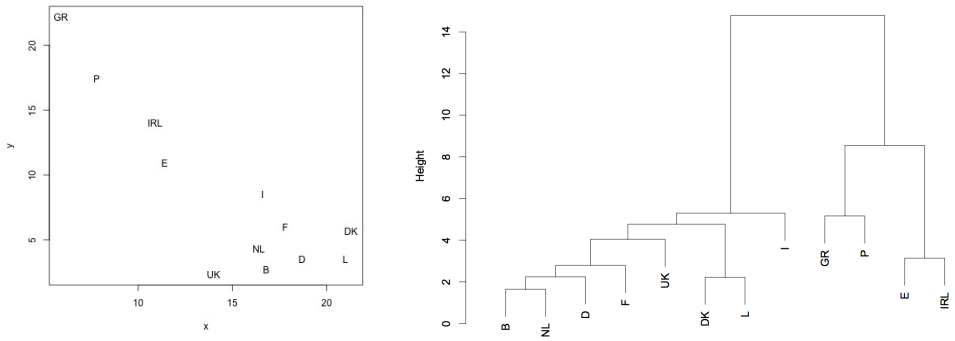

Figure 4: The agriculture dataset shown in the left panel and the resulting hierarchy created by the described algorithm (`diana`) shown in the right panel.

(dendogram) of the data, and consists of two types of methods. Agglomerative hierarchical clustering begins with each data point as its own small cluster, and merges until the entire dataset is part of a single cluster. Divisive hierarchical clustering starts with a single cluster and splits them until each point is in its own cluster. The clustering employed in this thesis is divisive, namely the function `diana` ([43] DIvisive ANAlysis) with the similarity measure defined by the Manhattan (Taxicab) metric.

Given a set of n points, the algorithm performs $n-1$ successive splits, where the cluster C with the largest diameter according to some metric d (in our case Manhattan), namely

$$\operatorname{diam}(C) = \max_{i,j \in C} d(i,j),$$

is selected to be split in each iteration. The following is pseudcode for the algorithm to perform the split, a variant of the algorithm specified in [44]:

1. Initialize $A = C$, and $B = \emptyset$
2. Move a single object from A to B.
   For every object $i \in A$, calculate the average dissimilarity to all other objects of A, namely

   $$a(i) = \frac{1}{|A|-1} \sum_{j \in A, j \neq i} d(i,j).$$

   Move the object with the highest dissimilarity, namely $m = \operatorname{argmax}_i a(i)$ to B. Explicitly

   $$A = A \setminus \{m\},\ B = B \cup \{m\}.$$

3. Move different objects (called the splinter group) from A to B.
   If $|A| = 1$, conclude.
   Otherwise, calculate $a(i)$ for all $i \in A$ and the average dissimilarity of i to all objects of B, namely

   $$d(i,B) = \frac{1}{|B|} \sum_{j \in C} d(i,j).$$

   Then select the object $h \in A$ such that

   $$a(h) - d(h,B) = \max_{i \in A}(a(i) - d(i,B)).$$

   If $a(h) - d(h,B) > 0$, then move h from A to B and repeat from 2.
   If $a(h) - d(h,B) \leq 0$, conclude.

The `diana` algorithm run on the agriculture data set (containing single points from 12 countries) is shown in Figure 4. Once the hierarchy has been formed, the optimal clusters and number of clusters k

is found by cutting the tree $k-1$ times and measuring the quality of the resulting clusters, measured using the gap statistic [45]. Given the data clustered into k clusters $C_1, ..., C_k$, define the sum of pairwise distances for all points in a cluster C_r to be

$$D_r := \sum_{i,j \in C_r} d(i,j).$$

Then define

$$W_k := \sum_{r=1}^{k} \frac{1}{2|C_r|} D_r.$$

The optimal number of clusters is the k maximizing the gap statistic $\text{Gap}_n(k)$, namely

$$\text{Gap}_n(k) = \mathbb{E}_n^*[\log(W_k)] - \log(W_k),$$

where $\mathbb{E}_n*$ denotes the sample expectation of a null reference distribution of the data.[3]

3 Analyses

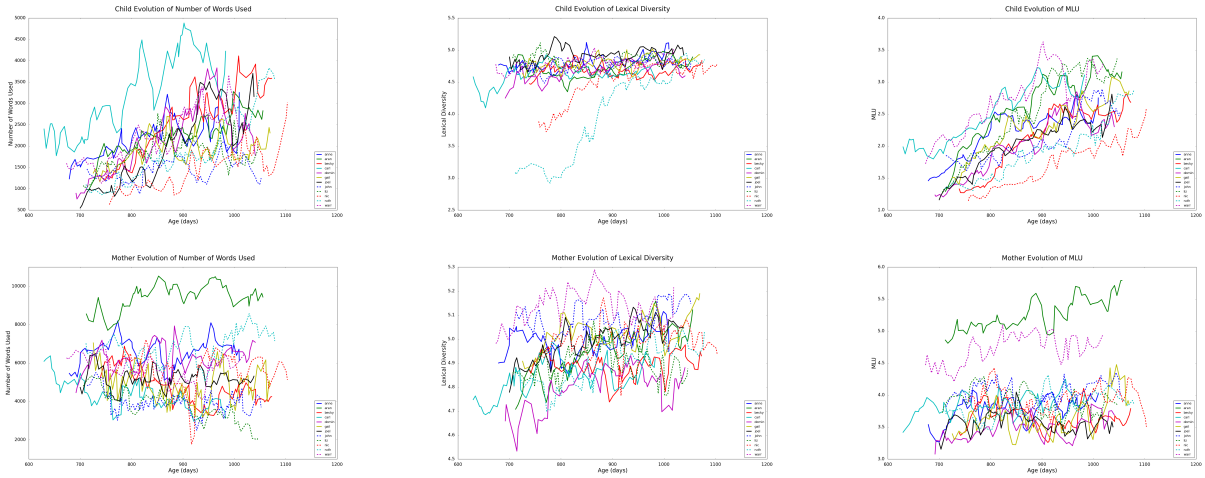

Figure 5: Evolution of the measures under consideration as a function of the children's ages, for the children (top row) and their mothers (bottom row). The total number of words produced are plotted in the left column, the lexical diversities in the middle column, and the MLUs in the right column.

The corpus used is the Manchester data set in the CHILDES database [47, 7], containing annotated transcripts of audio recordings from a longitudinal study of 12 British English-speaking children (6 girls and 6 boys) between the ages of approximately two and three years. The children were recorded at their homes for an hour while they engaged in normal play activities with their mothers. Each child was recorded on two separate occasions in every three-week period for one year. Each recording session is divided into two half-hour periods. The annotations include the lemmatized form of the words (lemma here means the canonical form of a word, for example "is" and "being" are considered instances of "be", "walking" and "walked" are considered instances of "walk") produced by both the children and their mothers (incomplete words and small word-internal errors were manually corrected in the lemmatization).[4]

[3] This clustering in the analysis was performed using the `cluster` R package [46].

[4] The CHILDES Corpus Reader, within the NLTK Python library, is used to read this XML formatted data and store them in more convenient structures for data analysis. The initial data munging can be found in the beginning of the code at the end of this thesis.

In order to increase the sample size in each period, a sliding window technique was used (akin to that used in [48]), by computing measures for the samples contained in overlapping windows of three consecutive corpus files. In this way, at each point obtained samples originating from two files from the same recording session were obtained, and a file from either the previous or the next recording session. For each child and mother, the total number of words they produced was recorded, the lexical diversity measured as the entropy of the lemmas produced (following the estimation method of [5], which is demonstrated to be accurate and unbiased for these sample sizes), and the mean length of the utterances (MLU) they produced. Instead of measuring MLU in morphemes [49], the simpler, but equally accurate measure in number of words [50] was used. In these ages, MLUs are well known to provide an accurate measure of the syntactic richness of the utterances produced [49], and in fact correlate almost perfectly with explicit measurements of grammatical diversity as discussed above[5]. A measure of inflectional diversity [48, 5] was also tested, which was not found to produce any reliable causal effects (7). However, the presence of these additional tests was nevertheless taken into account when correcting for multiple comparisons. Fig. 5 plots the temporal evolution of the three initial measures for the children and their mothers. Additionally, twelve more causal tests were conducted (namely cross-measurement analysis - mother MLU and child lexical diversity, etc. - see 8). These were again corrected for multiple comparisons, and these results can be found in the appendix.

The optimal time lag τ and embedding dimension E were estimated using the methods described earlier in the thesis. These values were found for each child time series and each corresponding mother time series. The twelve (τ, E) pairs (one for each child) were clustered using the `diana` package, and the optimal cluster number was found to be 1. Thus the estimates for the time lag and embedding dimension were not found to differ significantly across children or mothers, and therefore for each measure, a single estimate of (τ, E) for all children were used, and a single estimate for all mothers. The time series were checked to ensure that they contained non-linear signal not dominated by noise using a prediction test, and the presence of directional causality between children's and mothers' measures was tested for each of the three variables using mCCM, with 1,000 bootstrapping iterations used to assess the p-values.[5] Finally, to account for the lack of *a priori* predictions on the causal directions to be tested, the p-values were adjusted for multiple comparisons using the false discovery rate for correlated data (FDR; [42]).

4 Results

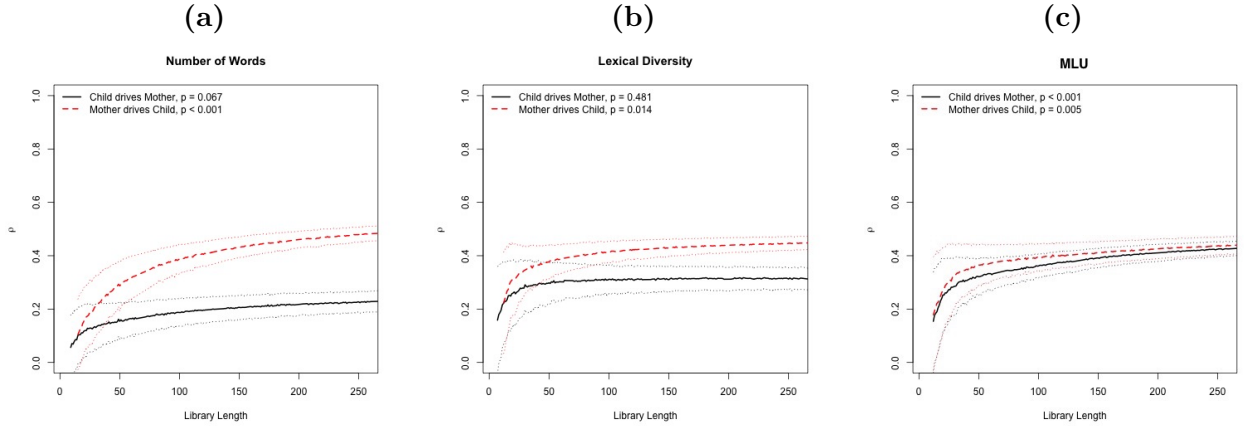

Figure 6: For each of the measures considered, the panels plot the evolution of Pearson's correlation coefficient (ρ) between the predicted and predicting shadow manifolds. The dashed lines denotes the standard deviation of the estimates. The p-values indicated in the legend were obtained by the bootstrapping procedure described by [9]. A value of ρ significantly increasing with library length is mCCM's indication of causality between two variables.

[5] All computations, except for τ selection, were done using R package `multispatialCCM` [9].

The three panels in Fig. 6 plot the results of mCCM for each pair of time series. Panel (a) shows that convergence indicates a significant causal relation between the number of words produced by the mother, and the number of words produced by her child ($p < .001$) which is not significantly present in the opposite direction ($p = .067$). A similar picture arises in the lexical diversities in panel (b): The richness of the vocabulary produced by the mothers influences the richness of the vocabulary produced by their children ($p = .014$), but the richness of the vocabulary used by children does not significantly affect that of their mothers ($p = .481$). In contrast, panel (c) shows that, in terms of MLU, the language produced by children and their mothers is strongly coupled, with significant causal relations in both directions (child to mother: $p < .001$; mother to child: $p = .005$).

5 Discussion

In terms of the amount of speech, or the richness of the vocabulary used, these results indicate that the mothers are not increasing the complexity of CDL in response to the details of CL, but the children's performance still benefits from the increased quantity and diversity of words. This is evidence that weak lexical fine-tuning serves a functional role. In contrast, when it comes to MLUs, the bidirectional causality provides clear evidence for a strongly coupled system with feedback. As is shown in [5], MLUs are in fact almost perfectly correlated (i.e., Pearson's $r \approx .96$) with an explicit measure of the diversity of the syntactic structures used in a sample (i.e., the *syntactic diversity* of the sample). Mothers adjust the complexity of their syntactic structures as a direct response to the syntactic complexity of the utterances produced by their children, as is advocated by the strong version of the fine-tuning hypothesis.

6 Conclusion and Future Work

The results provide direct evidence for the fine-tuning hypothesis. For the first time, it has been *explicitly* demonstrated that, in all measures studied, the children benefit from the gradual increase in complexity of CDL, as is indicated by the directional causalities found between the measures in CDL and those in CL. In addition, only for the case of syntax, we find direct evidence for the strong version of the fine-tuning hypothesis: The complexity of the syntactic structures produced by mothers are directly caused by those of the syntactic structures produced by their children. These findings are the first step in building a macroscopic level dynamical system model of language acquisition explicitly considering children jointly with their environment. In order to build such a model, one also needs to test for the explicit causal components of complexity within an individual (e.g., what are the causal connections between increased vocabulary and increased syntactic knowledge?), and those present across individuals and linguistic strata (see Appendix); for instance, it has been reported that the amount of speech produced by parents influences the growth of both the vocabulary (e.g., [51], [52]), and the MLUs in the children [53]. More in depth analysis of these initial results will be provided soon (see [54]) as well as the development of an intra-causal model of an individual's language complexity (see [55]). Moreover, we have begun analysis on the same causal relationships now using a Hebrew corpus, and hope to extend the analysis to several more languages.

7 Appendix

7.1 Additional Results

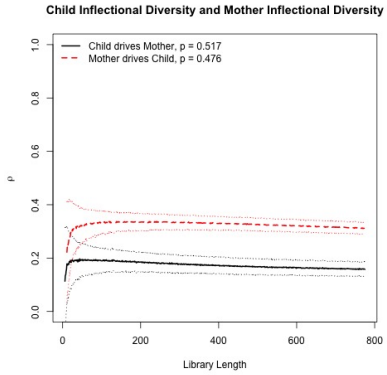

Figure 7: The panel shows that no significant causal relation was detected between the child and the mother for the inflectional diversity measure.

Fig. 7 shows no detected causal relation between the child's and the mother's inflectional complexity (child to mother: $p = 0.517$; mother to child: $p = 0.476$). However, the twelve panels in Fig. 8 show many significant causal relations (all panels except for (d)). The mother drives the child's language complexity in almost every case, and the child's syntax (MLU) is the only measure that contributes to the change in the mother's language complexity.

7.2 Source Code

Python data munging and time series measurements:
`https://github.com/jirvin16/nltk_childes/tree/master/English/multispatial_childes.bkr`
R state space reconstruction and causal analysis in English (including new analyses):
`https://github.com/jirvin16/nltk_childes/tree/master/English/total_analysis.R`
`https://github.com/jirvin16/nltk_childes/tree/master/English/plot_analysis.R` R state space reconstruction and causal analysis in Hebrew:
`https://github.com/jirvin16/nltk_childes/tree/master/Hebrew/total_analysis.R`
`https://github.com/jirvin16/nltk_childes/tree/master/Hebrew/plot_analysis.R`

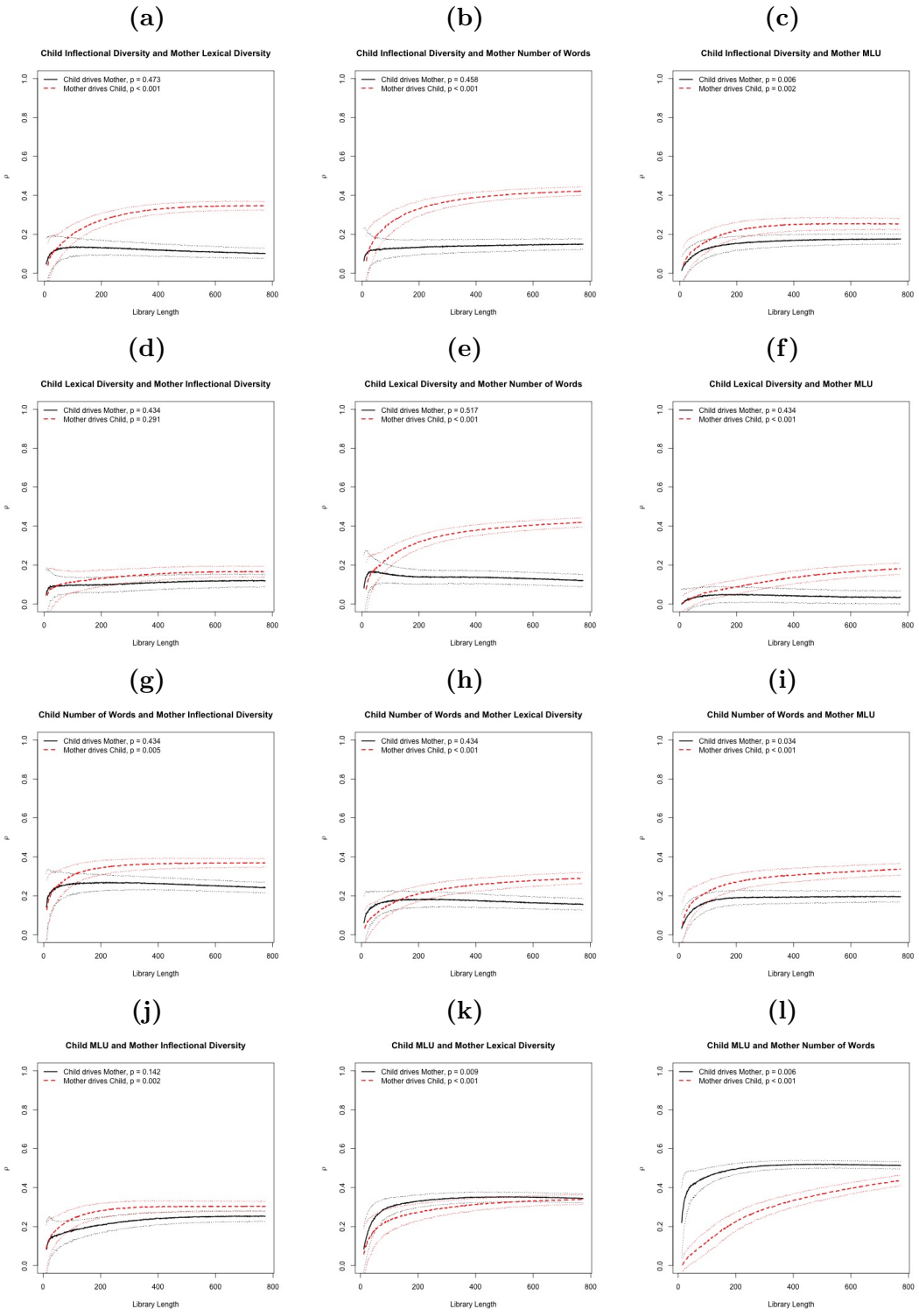

Figure 8: Again, for each of the pairs of measures considered, the panels plot the evolution of Pearson's correlation coefficient (ρ) between the predicted and predicting shadow manifolds. A value of ρ significantly increasing with library length is mCCM's indication of causality between two variables.

References

[1] Henry D. I. Abarbanel, Reggie Brown, John J. Sidorowich, and Lev Sh. Tsimring. The analysis of observed chaotic data in physical systems. *Reviews of Modern Physics*, 65:1331–1392, 1993.

[2] J. P. Huke. Embedding nonlinear dynamical systems: A guide to takens' theorem. Technical report, University of Manchester, 2006.

[3] Floris Takens. Detecting strange attractors in turbulence. In D. A. Rand and L.-S. Young, editors, *Dynamical Systems and Turbulence*, pages 366–381. Springer Verlag, Berlin, Germany, 1981.

[4] Michael Spivak. *A Comprehensive Introduction to Differential Geometry*, volume One. Publish or Perish, Inc, third edition, 1999.

[5] Fermín Moscoso del Prado Martín. Vocabulary, grammar, sex, and aging. *Cognitive Science*, in press.

[6] Paul L. C. van Geert. A dynamic systems model of cognitive and language growth. *Psychological Review*, 98:3–53, 1991.

[7] Brian MacWhinney. *The CHILDES Project: Tools for analyzing talk*, volume 2: The database. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition, 2000.

[8] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan R. Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *Science*, 338:496–500, 2012.

[9] Adam Thomas Clark, Hao Ye, Forest Isbell, Ethan R. Deyle, Jane Cowles, G. David Tilman, and George Sugihara. Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology*, 96:1174–1181, 2015.

[10] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[11] Catherine E. Snow and Charles A. Ferguson, editors. *Talking to Children*. Cambridge University Press, Cambridge, England, 1977.

[12] Toni G. Cross. Mothers' speech adjustments: The contribution of selected child listener variables. In Catherine E. Snow and Charles A. Ferguson, editors, *Talking to Children*, pages 151–188. Cambridge University Press, Cambridge, England, 1977.

[13] Catherine E. Snow. Understanding social interaction and language acquisition; sentences are not enough. In Marc H. Bornstein and Jerome S. Bruner, editors, *Interaction in human development*, Crosscurrents in contemporary psychology, pages 83–103. Lawrence Erlbaum Associates, Hillsdale, NJ, 1989.

[14] Jeffrey L. Elman. Learning and development in neural networks: the importance of starting small. *Cognition*, 48:71–99, 1993.

[15] Peter F. Dominey and Christelle Dodane. Indeterminacy in language acquisition: the role of child directed speech and joint attention. *Journal of Neurolinguistics*, 17:121–145, 2004.

[16] Steven Pinker. *The language instinct: How the mind creates language*. Harper-Collins, New York, NY, 1994.

[17] Catherine E. Snow. Issues in the study of input: Fine-tuning, universality, individual and developmental differences, and necessary causes. In Paul Fletcher and Brian MacWhinney, editors, *The Handbook of Child Language*, pages 180–193. Blackwell, Oxford, England, 1995.

[18] Jeffrey L. Sokolov. A local contingency analysis of the fine-tuning hypothesis. *Developmental Psychology*, 29:1008–1023, 1993.

[19] Elisa Newport, Henry Gleitman, and Lila Gleitman. Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In Catherine E. Snow and Charles A. Ferguson, editors, *Talking to Children*, pages 109–149. Cambridge University Press, Cambridge, England, 1977.

[20] H. Scarborough and J. Wycoff. Mother, I'd still do it myself: Some further non-effects of 'motherese'. *Journal of Child Language*, 13:431–437, 1986.

[21] Virginia Valian. Input and language acquisition. In William C. Ritchie and Taj K. Bhattia, editors, *Handbook of child language acquisition*, pages 497–530. Academic Press, San Diego, CA, 1999.

[22] Janellen Huttenlocher, Marina Vasilyeva, Elina Cymerman, and Susan Levine. Language input and child syntax. *Cognitive Psychology*, 45:337–374, 2002.

[23] Richard Kunert, Raquel Fernández, and Willem Zuidema. Adaptation in child-directed speech: Evidence from corpora. In *SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue*, pages 112–119, Los Angeles, CA, 2011.

[24] A. Murray, J. Johnson, and J. Peters. Fine-tuning of utterance length to preverbal infants: effects on later language development. *Journal of Child Language*, 17:511–525, 1990.

[25] Brandon C. Roy, Michael C. Frank, and Deb Roy. Exploring word learning in a high-density longitudinal corpus. In Niels Taatgen, Hedderik van Rijn, Kambert Schomaker, and John Nerbonne, editors, *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2106–2111, Austin, TX, 2009. Cognitive Science Society.

[26] Brian MacWhinney. What we have learned. *Journal of Child Language*, 41:124–131, 2014.

[27] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, pages 27, 379–423, 623–656, 1948.

[28] George A. Miller. Note on the bias of information estimates. In Henry Quastler, editor, *Information Theory in Psychology*, pages 95–100. Free Press, Glencoe, IL, 1955.

[29] J. F. Heltshe and N. E. Forrester. Estimating diversity using quadrat sampling. *Biometrics*, 39(4):1073–1076, Dec 1983.

[30] Anne Chao, Y. T. Wang, and Lou Jost. Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution*, pages 4, 1091–1100, 2013.

[31] Alexander Kraskov, Harald Stogbauer, and Peter Grassberger. Estimating mutual information. *Physics Review*, May 2003.

[32] Linda Smith and Esther Thelen. Development as a dynamic system. *TRENDS in Cognitive Sciences*, 7:343–348, 2003.

[33] Paul L. C. van Geert. *Dynamic Systems of Development: Change Between Order and Chaos*. Harvester Wheatsheaf, Hemel Hemstead, England, 1994.

[34] Rick Dale and Michael J. Spivey. Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, pages 391–430, 2006.

[35] Hederien W. Steenbeek and Paul L. C. van Geert. A theory and dynamic model of dyadic interaction: Concerns, appraisals, and contagiousness in a developmental context. *Develpmental Review*, 27:1–40, 2007.

[36] Antonio and Fabio Di Narzo. *tseriesChaos: Analysis of nonlinear time series*, 2013. R package version 0.1-13.

[37] Adam Clark. *multispatialCCM: Multispatial Convergent Cross Mapping*, 2014. R package version 1.0.

[38] Judea Pearl. *Causality: models, Reasoning and Inference*. Cambridge University Press, Cambridge, England, 2000.

[39] Clive W. J. Granger. Testing for causality. *Journal of Economic Dynamics and Control*, 2:329–352, 1981.

[40] Cunlu Zou and Jianfeng Feng. Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC Bioinformatics*, 10:122, 2009.

[41] Chih-hao Hsieh, Christian Anderson, and George Sugihara. Extending nonlinear analysis to short ecological time series. *The American Naturalist*, 171:71–80, 2008.

[42] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001.

[43] Anja Struyf, Mia Hubert, and Peter J. Rousseeuw. Clustering in an object-oriented environment. *Journal of Statistical Software*, 1996.

[44] P. Macnaughton-Smith, W.T. Williams, M.B. Dale, and L.G. Mockett. Dissimilarity analysis: a new technique of hierarchical sub-division. *Nature*, 202:1034–1035, 1964.

[45] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society*, 63:411–423, 2001.

[46] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2015. R package version 2.0.3.

[47] A. L. Theakston, Elena V. M. Lieven, Julian M. Pine, and C. F. Rowland. The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, 28:127–152, 2001.

[48] Ferm´ín Moscoso del Prado Martín. Grammatical change begins within the word: Causal modeling of the co-evolution of Icelandic morphology and syntax. In P. Bello, M. Guarini, M. McShane, and B. Scasselatti, editors, *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 2657–2662, Austin, TX, 2014. Cognitive Science Society.

[49] R. Brown. *A first language: the early stages*. Harvard University Press, Cambridge, MA., 1973.

[50] Matthew D. Parker and Kent Brorson. A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language*, 25:365–376, 2005.

[51] Nereyda Hurtado, Virginia A. Marchman, and Anne Fernald. Does input influence uptake? links between maternal talk processing speed and vocabulary size in Spanish-learning children. *Developmental Science*, 11:F31–F39, 2008.

[52] Adriana Weisleder and Anne Fernald. Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, 24:2143–2152, 2013.

[53] Sally Barnes, Mary Gutfreund, David Satterly, and Gordon Wells. Characteristics of adult speech which predict children's language development. *Journal of Child Language*, 10:65–84, 1983.

[54] Jeremy Irvin, Daniel Spokoyny, and Fermín Moscoso del Prado Martín. Dynamical systems modeling of the childmother dyad: Causality between child-directed language complexity and language development. In Anna Papafragou, John Trueswell, Dan Grodner, and Dan Mirman, editors, *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Austin, TX, 2016. Cognitive Science Society.

[55] Daniel Spokoyny, Jeremy Irvin, and Fermín Moscoso del Prado Martín. Explicit causal connections between the acquisition of linguistic tiers: Evidence from dynamical systems modeling. *Association for Computational Linguistics*, in press.