Evolution of compositional languages in multiple agent social communities

by

Shashank Srivastava

(Y5827429)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

August 2010

Evolution of compositional languages in multiple agent social communities

A Thesis Submitted

in Partial Fulfillment of the Requirements

for the Degree of

Master of Technology

by

Shashank Srivastava

to the

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

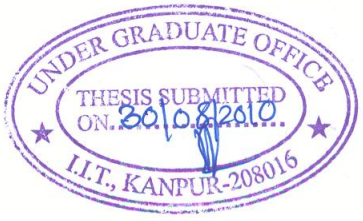
August 2010

CERTIFICATE

It is certified that the work contained in the thesis entitled **Evolution of compositional grammars in multiple agent social communities**, by **Shashank Srivastava**, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

August 2010

(Harish Karnick)
Professor,
Department of Computer Science & Engg,
Indian Institute of Technology,
Kanpur-208016.



Dedicated to my loving Grandparents.

Acknowledgements

I would like to thank Dr. Harish Karnick for guiding me throughout my thesis work and also otherwise, providing valuable insights whenever I was stuck, for letting me work at my own pace; and for being a being a most kind person.

I also wish to thank my friends who made my stay at this institution enjoyable.

Shashank Srivastava

Abstract

While computational modeling has yielded several plausible models for language emergence in a set of uniformly endowed agents, most of these treatments do not address emergence of syntax; and have focused on the evolution of a coherent lexicon. While a coherent vocabulary is a necessity for any language, it is in fact syntax which allows humans to express seemingly infinite meanings using a finite set of phonetic elements. Most earlier models also ignore population turnover, and do not incorporate dynamical and structural aspects of populations.

In this thesis, we have extended a well known inductive learning model of language learning to large populations, heteregeneous interactions, and realistic social communitites. The model induces grammatical rules on the basis of phonetic resemblances between lexical entities, and similarities in semantic meanings they correspond to. We have developed a framework where multiple agents can interact in an iterated learning setting, and each agent can receive its primary linguistic input from a set of speakers according to distributions specified by the existing social topology. We also try to extend the deterministic production model to a probabilistic one, and investigate possible biases which can expedite the emergence of compositional syntax.

In particular we study the effect of population size and the structure of social topology on linguistic coherence and language emergence for this model. Our investigation of the extended model on different social graphs leads to several insights, and indicate that social topology can have significant effects on the acquisition and evolution of language.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Language is the means of getting an idea from my brain into yours without surgery."
-Mark Amidon

Questions of the emergence and evolution of human language are among fundamental questions in social and cognitive sciences. Human language exhibits a range of unique features such as semantic structure, open-endedness and compositionality; unmatched elsewhere in the natural world. While a few other species exhibit a limited lexicon (in primates), linguistic pattern learning (in primates) [4], and rudimentary syntacic structure (in the dance of bees) [3], no other species exhibits linguistic behaviour resembling the complexity of human language[5, 6]. It is in fact commonly recognized that the complexity of human thought, development of mental concepts and cognitive awareness are inseparably connected with, and intimately entwined with human language development.

A major issue in the study of human language evolution is a lack of historical data. The development of languages are one-time epochal events, with no stratified records at various phases of their development. Empirical clues about language acquisition and emergence only come from the study of contemporary pidgins and creoles, or the study of language acquisition in children.

1.1 The Nativist Approach

Presenting the precocious use of language by human infants as evidence, the Chomskyian school of linguistics has proposed that all human languages share some linguistic primitives that render them easy to learn given very little exposure (argument of 'poverty of stimulus')[7][8]. The strong Chomskyian view holds that these linguistic primitives have a genetic basis, called a Language Acquisition Device(LAD); and that the major part of human language is hardwired into human brain structure. It claims that the basis of language, not unlike the visual apparatus [9], is in the expression of genes.

In this view, language is not acquired from the environment, but rather reached *internally* through a process of maturation. This treatment resolves the poverty of stimulus dilemma at the design level itself, suggesting that most of the syntactic structure of language is biologically specified. The role of the environment is minimal, and only foolows an internally pre-determined course of linguistic development. Indeed, the strong Chomskyian school holds that the LAD is an *organ*, no less real than the brain or the liver. Apart from poverty of stimulus, there are other factors, such as the emergence of syntactic language in a group of deaf Nicaraguan children [26], and the existence of a critical period for first language acquisition that are cited in support of the nativist view.

However, the Chomskyian perspective has several critics, especially because there is no plausible account of how the LAD evolved. Consequently, the view is far from being universally accepted.

1.2 And dissenters

The assumption that language has a purely or overwhelmingly genetic basis is not conclusively supported by empirical neurobiological studies. The most significant counter-argument against Chomskyian nativism's biological determinism is the variety in the structures of existing natural languages. The alternate empiricist view has suggested that rather than being an innate biological quality, the structure of language is *induc-*

tively inferred by children from primary linguistic input (PLI); and that the structure of language itself contains the seed for its learning.

While the strong empiricist view emphasizes the role of PLI and induction learning as sole determinants of a listener's linguistic hypothesis, and precludes any innate bias altogether, Deacon suggested a less extreme alternative. He suggested that rather than biological evolution being responsible to develop a LAD over time, it is languages themselves that stucture to adapt to some inherent biases in our cognitive structures, and limitations such as limited number of examples. This can explain the emergence of 'language universals' as well as linguistic divergence since the cognitive apparatus is almost the same, but linguistic evolution can follow widely different pathways.

The generic induction-learning argument has support from several observations in child language acquisition. Evidence suggests that linguistic patterns of children are very closely synchronous to that of their parents. Pinker [22] observes that children generalize a rule after sufficient number of exposures to instantiations of the same. Thus an agent can acquire prevalent linguistic behaviour purely based on observing relative frequencies of behaviours, and any inherent prewiring is not needed to explain the phenomenon. An Occam's razor view of the situation would suggest that a non-nativist hypothesis be preferred, that doesnt require additional support from an independent language acquisition organ.

Also, Child Directed Speech(CDS) or baby-talk universally exhibits modifications in linguistic behaviour such as smaller vocabulary and simplistic syntactic forms that might aid learning of a child. In some sense, this negates the argument of poverty of stimulus, by claiming that learnable linguistic stimulus is not as sparse as is often perceived.

1.2.1 Perspective from Learning Theory

The nativist school though, in some sense, can claim some mathematical basis from results in classical learning theory.

In 1967, Gold [11] laid a framework for language learnability as identifying a language in the limit of providing all strings in the language. Gold proved that no learning

algorithm A_H can learn from any set of super-finite languages (a set having all finite languages and at least one infinite language), under this framework. This argument however, is open to criticism on grounds that the notion of learnability is unnaturally restrictive, since children dont need to learn a language *exactly*. In fact, it would be hard to find two observers who would agree on the validity of *all* strings for any language. Indeed, linguistic mutations and the dynamic open-ended nature of human languages suggest that we learn languages imperfectly; and our linguistic hypothesis significantly differ from those of our cultural forebears.

Vapnik's framework of statistical learning theory [12] relaxed the constraint of perfect learning using the PAC learning model. The PAC framework essentially aims for a bound on a high probability of accuracy, while allowing a small probability of generalization error ϵ). This relaxation extends the class of learnable languages, but critically it reaffirms Gold's conclusion that learnability needs a constrained set of candidate hypothesis languages. This could be viewed as evidence in favor of the Chomskyian concept of *Universal Grammar*, in the sense that the learner needs an apriori bias about the hypothesis space, that assists the language acquisition process.

In summary, the dual fact that only humans can produce or learn languages; and that no learning algorithm A_H can be guaranteed to work well on unrestricted search spaces or all sorts of training data suggests that language needs at least some inherited subliminal biological/cognitive biases. On the other hand, it is unfounded and hard to argue that biological evolution has lead to a highly constrained and refined 'language organ', as championed by proponents of the Chomskyian school and Universal Grammar. If language acquisition indeed proceeds by, or is even marginally facilitated by a process of inductive learning, this implies that the external environment influences the structure and pedagogy of language. The questions to raise, then, are the extent to which language is learnt by inductive generalizations, the form and drivers of evolutionary linguistic biases, and the mutual roles of cultural adaptation and the foresaid biological evolution.

1.3 Complex Adaptive Approach

In this thesis we explore the possibility, as just suggested and first proposed by linguists including Briscoe and Steels [14][13], of treating language as an evolutionary phenomenon. This view of language treats it as a living 'complex adaptive system', that is continually adapting through co-evolving cultural and biological forces.

Language, in this sense, needs to be seen as a continually changing complex system (similar to an economy or an ecosystem) where local interactions lead to global organization, and the onus of adaptation is on language itself to be transmissible; within the constraints of available cognitive architecture, limited communication, and noisy signal perception. This is the sense in which the system is *complex*: there is no hierarchy or modular structure in the interactions, and global level properties are not obvious from the nature of local interactions. In this sense, features such as compositionality are seen as emergent(rather than predetermined) solutions in response to the system constraints and initial conditions.

The system evolves through interactions of the agents among themselves and with the environment, changing the state of each agent, as well as potentially both the environment and the cultural and social systems. The system is called *adaptive* since the future state of the system (consisting of states of individual agents and the environment) at any point is a probabilistic function of the current state.

In this work, we assume a population of artificial agents endowed with identical cognitive architecture, and investigate, as have others, how many features of natural languages can emerge from a richer treatment of language as a complex adaptive system, as evolutionary artefacts in a process of cultural evolution and social transmission in social communities over many generations.

1.4 Role of culture,population and topology

Social structuring started 250000 years ago when Homo hydelbergensis moved out of Africa. There is evidence of cave paintings, which possibly form the first signs of human

language development. It has been claimed that social topology, geographical locality and social behaviour are major determinants of linguistic evolution [16]. Most language models however ignore this structure, assuming uniform homogenous interactions at all population levels. Significantly, Nettle [17] shows that population structure can determine the problem of whether a rare linguistic mutation will propagate in a population. Ke [19] extended Nettle's results to more realistic networks, and showed that socially influential agents can affect the spread of a linguistic variation. Mague [25] applies Mufwene's basic model [24] of lexicon learning to multiple population structures, and shows dynamics of language learning are directly influenced by population structure. Similar results are reported by Lee et al [23], showing that an impoversished treatment of population structure in language models can lead to flawed and spurious generalizations.

1.5 Overview

Many models have shown and vindicated the emergence of stable vocabularies in populations of agents. The amalgamation of linguistic lineal units through social interactions across mutiple generations can lead to formation of stable coherent lexicons that form the heart of any human language. While computational modeling has yielded several plausible models for language emergence in a set of uniformly endowed agents, most of these treatments suffer from a lack of treatment of syntax [33], ignoring population turnover [30], and lack of population structure [1] or unrealistic population dynamics [31, 17, 19].

In this thesis, we seek to rectify some of these shortcomings by a more realistic treatment of language, and social structures that more closely resemble human interaction networks, at least in their rudimentary forms. In the process, we extend a well known Iterated Learning model by Kirby by introducing necessary biases, while highlighting some limitations of the model, especially in scenarios of multiple teachers and coexisting hypotheses.

In chapter 2, we briefly review major seminary works in language evolution, especially those following a computational modeling approach. In chapter 3, we explain the basic framework of our model, basic types of rule subsumption; and the application of the model to a population of agents. chapter 4 explores various population models and their emergent properties. In chapter 5, we explain the results, and conclude with a brief discussion and possible future directions of enquiry.

Chapter 2

Review of computational simulation models

As previously mentioned, the critical problem in studying language emergence and evolution is the lack of documented data. Most extant human languages emerged a long time ago, and new languages do not emerge frequently. Moreover, the evolution of languages is a gradual process, taking tens or hundreds of years. To conclude the linguist's litany of woes, the adaptation of cognitive architecture for language development in humans was a one time event. This alone makes questions about language evolution naturally hard to answer. Even fundamental issues such as emergence of features such as compositional syntax, and any analysis of evolutionary benefits of such adaptations are at best speculative.

Because of paucity of data, abstract theorizing and phylogenetic speculation was the *modus operandi* of research in this generic domain for a long time. With the advent of computational linguistics however, a host of analytical and simulation based techniques have been used to study language origin and evolution. Most analytical methods in linguistics can be characterized by macroscopic population analysis by researchers such as Niyogi [29], and function theoretic mathematical techniques as in work by Nowak and Komarova [27, 28]. Such works can predict large scale properties, and are mathematically

tractable.

An alternative approach to rigorous mathematical analyis is computational simulation modeling. Simulation models provide an intermediate way between abstract theorizing on one hand, and grounded rigorous mathematical bases on the other. They provide for intuitive modifications in the simulation setting, without needing to mathematically analyze them precisely. By a choice of formalism, the linguist can choose what aspects of language to study in isolation. A typical simulation model normally consists of an alternate world, where a set of agents with predefined capabilities can interact among themselves and with the environment, to evolve their linguistic behaviour. The choice of agents who interact is dictated by a defined policy but is otherwise random. For example, if agent groups are represented as the nodes of a graph then interaction may be restricted to agents of only directly connected nodes. In particular simulation models are the natural way to investigate complex and mathematically intractable systems, exhibiting non-linear behaviour.

In this chapter, we review briefly some relevant prior work in simulation modeling, based on evolution of lexicons and syntax.

2.1 Luc Steels

Luc Steels pioneered robotic simulations in evolutionary linguistics in the well known *Talking heads* experiment. The simulations [32, 33] consisted of robotic agents with mounted cameras, negotiating a virtual world where objects had a discrete set of features. Initially, agents had no innate semantic representation or lexical knowledge. However, they are endowed with memory structures for storing lexical entities, as well as hardwired procedures for parsing, or guessing the meaning of an unknown word. The agents are also endowed with consistent sensory channels, which can each faithfully transmit the value of one of several attributes for any perceived object.

The agents intially played *discrimination games* to identify an object by building decision trees based on the object's sensory features. Next, agents participate in interactions

called *guessing games*. A guessing game consists of a listener and a speaker randomly chosen from the population. The listener and the speaker are co-located and share the same context. The speaker selects one object from the context (called 'topic'), and also attributes that distinguish the topic from the rest of the context. If current distinctions are not sufficient to distinguish the object from the background, new distinctions can be created in an agent's internal decision trees. The speaker verbalizes the attribute through its existing lexicon, after which the listener tries to identify the attribute, and hence the topic using it's own lexicon and decision trees. The game is considered a success if the listener successfully identifies the topic. Otherwise, the speaker identifies the correct topic by a joint attention mechanism. Both the speaker and the listener update their memories according to the outcome of the game. The update may involve a new insertion into the lexicon, or a modification of the meaning-signal association matrix. The representation of the lexicon as an association matrix allowed for situations of both polysemy and homonymy. Steels' system preferred more successful meaning-signal pairs through a positive feedback between usage of a word in a word game, and its success. Steels' experiments showed that eventually all agents develop roughly the same discrimination trees, and hence the same object categories. Additionally, a joint lexicon emerged in the population of agents.

The crux of the experiments was that a set of artificial agents could develop common perceptual distinctions and a coherent lexicon, without any innate or biologically transmitted linguistic information. Also, this underplayed the significance of intergenerational cultural transmission in developing a coherent communication system. Instead, it showed that common cognitive abilities can suffice to evolve a successful communicative system in a population of agents.

2.2 A. Smith

Smith [36] used Steels' *talking heads* framework to show that lexical coherence can develop even in the absence of communication of explicit linguistic success. In this scenario,

unlike in the Talking Heads' experiments, agents play selfish language games, i.e. the listener never gets to know the intended meaning of the speaker. Thus there is no positive feedback to reinforce more successful meaning-to-signal mappings. Instead, meaning is inferred from the perceived context. In this case, it was shown that multiple exposures to the listener's language in different contexts, and comparision of overlap in the contexts can be used to infer the intended meanings.

This work was inspired by language learning in children, where explicit meaning is rarely communicated but children are exposed to multiple mappings for similar objects. They manage to acquire the correct mappings from existing linguistic hypotheses, presumably by picking common elements of shared mappings.

Thus, Smith's work addresses the poverty of stimulus problem, since it more realistically models how language acquisition occurs in humans, where in most communication there is no overt transference of intended meaning.

2.3 Ted Briscoe

Briscoe [14, 15] was among the first to view language as a complex adaptive system. His approach is also among the relatiively few which have tried to go beyond the lexicon and tried to incorporate elementary syntax into simulations, a complex and yet characteristic feature of human language. In his approach, the language of an agent as defined by a set of 20 parameter values (which can be viewed as its linguistic apparatus, or LAD), which largely focus on word order.

Each agent is modeled with a parser and an inbuilt parameter setting algorithm, which changes the values of the parameters based on observed language triggers. The parser works by modifying the contents of a stack containing word categories corresponding to the input sentence. Agents are also endowed with an innate parameter setting algorithm. If the parser fails on an input sentence, the parameter setting algorithm resets the values of one of the parameters. A parameter is reset just once in the process of linguistic acquisition by the agent. The choice of parameter to be reset is based on its

position in the hierarchy of parameters (most general parameters are reset first).

Briscoe's contribution was to show that a parameter-setting approach can lead to linguistic convergence at a relatively high level (syntactic). In his model, biological transmission does take place, and linguistic parameters get set to fixed values as we move through the simulation (for example, the word order parameter can be fixed to a particular word order). This can be seen as evolution of the linguistic apparatus. In this way, the model shows co-occurring adaptation of both the LAD and language. Linguistic convergence was achieved much faster when the initial parameters matched the language triggers agents are exposed to in the learning phase. Thus, even a simplistic LAD could be seen to evolve priors which would make language transmission and acquisition more effective.

However, there are obvious issues in these claims. In Briscoe's model, linguistic evolution proceeds at an unrealistic pace when compared with the phylogentic timescale of evolution. In reality, biological evolution is slower by several orders of magnitude. Large scale genetic modifications only become manifest over periods of thousands of years, while language forms can transform within a matter of a few decades or centuries. In this way, linguistic embeddings at the level of genes would always be 'moving targets' for biological evolution, as shown in simulations by Chater et al [39].

2.4 Christiansen and Devlin

Christiansen and Devlin's [38] simulations demonstrated that syntactic structures of language might have evolved to fit human sequential learning mechanisms. They claimed that constraints on the human learnability of sequential structure have steered the course of syntax evolution in languages; and are reflected as word order universals.

Their simulations consisted of training simple recurrent networks (SRN's) on 32 different grammars. Chosen grammars had varying levels of head-order consistency, i.e. different ratios of head-first and head last phrases. The task for SRN's was to predict the category of the next lexical entitty in a sentence. It was found that there was a large

correlation between head order consistency of grammars, and the accuracy of SRN's in learning them. Head-order inconsistent grammars were observed to be harder to learn. Further analysis of human languages revealed that languages containing patterns that SRN's found harder to learn were much rarer.

Similar results were later reported by Lupyan and Christiansen [40] by a frequency analysis of case-markings in human languages, and related the learnability of languages in artificial networks with the frequencies of occurrence of different word order forms, consisting of subject (S), verb (V) and object (O). Subject-first languages (SVO and SOV), which dominate human languages (74%) were found to be easiest to learn by the networks. OVS and OSV forms, which have very rare occurrence in world languages, were not easily learned.

Further, Conway and Christiansen [37] reviewed sequential learning in non-human primates, and found that learning behaviour for short length sequences and certain patterns were largely similar to humans. However, other primates are unable to deal with hierarchical sequential structure that is characteristic of human languages. This could partly explain why humans alone have complex syntactic linguistic abilities.

These simulations and studies together suggest a plausible alternative explanation of word-order emergence that avoids intractables such as Universal Grammar. They also stongly argue that languages, guided by capabilities and limitations of human learning and processing, primarily adapt culturally; while undercutting the influence of biological evolution or a Chomskyian LAD in governing syntactic features of language.

2.5 Zuidema

Zuidema [49] argues that language acquisition in an iterated learning setting is a special type of learning problem, since in this case the outcome of a learning process becomes the target of another learning process. He shows that because of the special nature of this learning process, language itself can adapt to the learning algorithm. Zuidema's language acquisition algorithm consists of three stepwise operations: (i) incorporation,

which involves extending the extant language to include the current string, (ii) compression, involving substituting frequent and long substrings of signals with non-terminals, and (iii) generalization, involving merging of two non-terminals. He demonstrates that in this learning model, early learners are unable to learn the target language. However, after several generations , an imposed compression based prior for CFG's leads to emergence of language which are reliably learnt by his learning procedure. This contradicts Nowak's argument of a 'coherence threshold', which is the minimum limit of the learning accuracy of an individual that could be consistent with global linguistic coherence in a majority of a population.

While Zuidema's compression bias undermines his claim of the model having learnt 'unlearnable grammars' (as per Gold's results), he effectively demonstrates how languages can emerge to fit biases in the learning procedures.

2.6 Simon Kirby

Kirby [1] demonstrated that complex properties of natural language such as compositional syntax and recursion can emerge in a synthetic framework over generations of interactions. Kirby claims that these properties must "inevitably emerge through the complex dynamical process of social transmission", and that transmission bottlenecks serve as drivers for such transformations.

Kirby's framework consists of an explicitly structured predicate-argument meaning space, and a greedy rule subsumption algorithm. The simulation follows an Iterated Learning Model [45, 20] of interaction, with the agent in Generation i always in the role of a listener, which subsumes rules based on linguistic input from its parent in Generation $i-1$. After a fixed number of iterations, the Agent $i-1$ 'dies', the Agent i takes over the role of the speaker and a new Agent $i+1$ with a blank grammar is introduced in the role of a new listener. The major claim is that some languages will be more easily transmissible through *transmission bottlenecks*, and these are the languages which will persist. This was reiterated in a different work by Brighton [10], showing that compositional languages

have a stability advantage over holistic languages for structured meaning spaces with few training examples by comparing a wide range of meaning spaces, and other parameters.

The emergence of compostitionality and recursion however, seem to be functions of the formulation and the learning algorithm, and not inevitable results of social transmission and adaptation as Kirby claims. The induction algorithm overtly favors smaller, compositional grammars. The kernel of the approach however, as far as we are concerned, is that it realistically shows languages adapting and evolving, emergence of competing word orders, and the rather interesting emergence of word categories, such as nouns, verbs and fillers, corresponding to semantic categories in the internal representations.

2.7 Henry Brighton

Brighton [34] addressed the problem of choosing from a set of hypothesis, that forms the core of the induction learning approach. In any agent based model, a linguistic agent needs to choose from a set of possible hypotheses $\mathcal{H}$, on exposure to training data $\mathcal{D}$. The chosen hypothesis should ideally be concise, and not be overly complex and hard to learn. At the same time, it should sufficiently and correctly explain the training data. Brighton formulated this tradeoff between simplicity and precision as a Minimum Description Length (MDL) problem, and validated his approach using a transducer based learning approach. The MDL principle provides a way of choosing, given the set of hypothesis $\mathcal{H}$ and data $\mathcal{D}$, which member of the hypothesis set represents the most probable hypothesis, given that the data $\mathcal{D}$ was observed.

Brighton's claim was that external constraints like MDL can explain universal properties such as compositionality and there is no need to invoke a biological LAD.

Chapter 3

Approach

In this chapter, we describe the experimental setup and the learning model we use in our study. Our model is motivated by the one proposed by Kirby [1] as discussed in the previous chapter. We seek to extend the Iterated Learning Model proposed by Kirby to more realistic population sizes and heterogeneous interactions between agents. The cognitive capacities of agents and the nature of the meaning space are identical to those in Kirby's study. The extension of the model from a few agents to a large population scenario presents several difficulties. We also attempt to generalize the model to simulate probabilistic linguistic behaviour by agents holding multiple hypothesis at the same time, and identify problems with this approach.

3.1 Basic setup

The Kirby world consists agents equipped with identical sets of meanings, an identical internal representation of language, and an induction algorithm that can move towards more compositional grammars. There is an explicit and globally homogenous I-language, where atomic concepts such as John, Gavin, Mary, likes, or knows can combine as predicate-argument propositions to form more complex meanings such as *loves(Mary, John)* or *knows(Gavin,loves(Mary,John))*.

The agents also have the ability to represent language through internally stored pro-

duction rules, which can generate Context Free languages. While the linguistic apparatus allows for features such as compositionality and recursion, they are not necessitated by it. For instance the meaning *kills(John,Mary)* can be expressed in the setup with the same signal '*johnkillsmary*', but using multiple possible grammars.

Grammar 1:

S/kills(John, Mary) $\longrightarrow$ *johnkillsmary*

Grammar 2:

S/1(John, 2) $\longrightarrow$ *john*A/1B/2

A/1 $\longrightarrow$ *kills*

B/2 $\longrightarrow$ *mary*

Grammar 3:

S/1(2, 3) $\longrightarrow$ M/2L/1N/3

L/kills $\longrightarrow$ *kills*

M/John $\longrightarrow$ *john*

N/Mary $\longrightarrow$ *mary*

In the above scheme, the left hand side of a rule is a semantic representation, possibly only partially specified, with capital letters signifying learned word classes. The numeric variables represent pre-terminals which can expand to fully specified meanings through other rules of production. The right hand sides represent (partially specified) lexical utterances, which can be construed to be sequences of atomic phonemes that the agents can produce and distinguish. We assume that the vocal and auditory apparatus is identically developed for all agents in the simulations, and hence this inventory is the same. The bigger assumption, seemingly, is on the consistency of the internal representations of the left hand side. However, Steels' [33] observation of the formation of coherent discrimination trees and general consistency of perception in humans suggests that even

if the internal representations are distinct, their net effect and workings are reasonably consistent.

In the first grammar, the signal-to-meaning mapping is completely arbitrary, and the language is holistic. No subpart of the holistic signal conveys any meaning. While decoding, the string is not chunked into smaller parts, and the whole string needs to be processed as a holistic entity. On the other hand, the third grammar is perfectly compositional in the sense that meaning conveyed by a string is a function of its parts. While the setup allows for such syntax to be functional, it is not initially hard-coded into the system, i.e. the grammars are not explicitly needed to be compositional. In fact, new individuals always begin with a blank grammar, and initially language consists of just holistic rules of production.

3.2 Language games

Individuals start as 'listeners', and subsequently graduate to become 'speakers'. Under the Iterated Learning Model, an agent in generation i receives its primary linguistic input from its linguistic parent in generation $i-1$, chooses its linguistic hypothesis through an induction algorithm, and then becomes a speaker to provide linguistic input to generation $i+1$.

3.2.1 Production mechanism

As a speaker, individuals need to convey randomly chosen meanings through their learnt production rules. If an agent can produce the desired meaning from its current linguistic hypothesis, it does. In case the current language cannot produce the meaning, it produces the signal (string) corresponding to the closest meaning that it can produce, and replaces the non-matching part with a randomly invented string of phonemes. For example, suppose an agent currently has the following grammar:

Grammar 4:

S/fly(Rhinos, 1) $\longrightarrow$ *rhinoscanflyA/1*

A/Aeroplanes $\longrightarrow$ *aeroplanes*

A/Zeppelins $\longrightarrow$ *zeppelins*

If the agent needs to convey 'fly(Rhinos,Helicopters)', it chooses a closest meaning (say 'fly(Rhinos,Aeroplanes)') that it can convey, and replaces the non-matching part of the parse-tree (that produces the string *aeroplanes* from the semantic meaning 'A/Aeroplanes') with a random string of phonemes [1], say *wxyz*. Thus the utterance for 'fly(Rhinos,Helicopters)' becomes *rhinoscanflywxyz*. On the other hand, suppose the agent has to convey the proposition 'fly(Hippos,Aeroplanes)', there is no non-trivial parse tree possible in the existing grammar that can be consistent with any of the three atomic meanings, and hence an entirely new parse tree needs to be constructed corresponding to an entirely random phonetic sequence, say *'jdix'*. Thus the utterance corresponding to 'fly(Hippos,Aeroplanes)' is the entirely non-compositional '*jdix*. [2]

The production mechanism for the speaker, therefore, needs two procedures: one is to deterministically check if it is possible to produce a meaning from a given grammar and generate the corresponding utterance. A recursive pseudocode for the procedure is given below:

```
PROCEDURE DEFINITION:
proc CanIProduce(Target_meaning, Current_symbol){
boolean=0;          //Can grammar G produce the Target_meaning?

for i=1:Number of production rules in Grammar G,
{
    C=Category symbol of rule i;
    S=Partial semantic meaning on LHS of rule i;
```

[1] In this work, the random strings are chosen with a maximum length of three, and there can be one of 15 phonemes at each position, represented by the first fifteen English alphabet

[2] This does not reflect an inadequacy in the generalization rules. If 'fly(Rhinos,1)' is an often reinforced semantic structure, while the flying of anything else(including hippos) is a rare observation, it is not entirely unnatural for 'fly(Rhinos,1)' to be learnt as a predicate with a single argument.

```
    RHS=Partial signal on RHS of rule i;

    //For a rule with matching symbol:
    if(C ~= Current_symbol) continue; end

    Replace variables in 'S' with corresponding category symbols from 'RHS';

    [b(1):b(k)]= Atomic elements or category symbols in the semantic S.
    // which should not be more than that in Target.
    [a(1):a(k)]= Corresponding k elements in the 'Target', determined by paranthesis.

    flag=1;
    //Match all k tokens:
    for j=1:k {
        if(b(j) is an atomic meaning){
            if(a(j) ~= b(j)) flag=0; end; }

        else if (b(j) is a category symbol){
            boolean2= CanIProduce(a(j),b(j));
            if(~boolean2) flag=0; end}
       }

    if(flag==1) return 1; end;
}
return 0;

CALLING THE PROCEDURE:
bool= CanIProduce(Target_meaning, 'S');
//'S' is start symbol in the grammar.
```

3.2.2 Invention procedure

A second procedure is needed to formalize the notion of similarity of representations of meanings in the I-space, for the invention mechanism to work on, in case the exact meaning cannot be produced by a grammar. This must take into account not just the structure of the semantic meanings, but their pathways of production (as seen earlier in this section). For instance, consider a hypothetical Grammar 5, an extension of our previous grammar.

Grammar 5:

S/fly(Rhinos, 1) $\longrightarrow$ *rhinoscanflyA/1*

A/Aeroplanes $\longrightarrow$ *aeroplanes*

A/Zeppelins $\longrightarrow$ *zeppelins*

S/know(1,2) $\longrightarrow$ B/1C/2

B/Rhinos $\longrightarrow$ *rhinonoun*

B/Hippos $\longrightarrow$ *hippob*

C/Rhinos $\longrightarrow$ *hippoc*

C/fly(Rhinos,Helicopters) $\longrightarrow$ *abc*

Suppose, we again needed to convey the unlikely proposition 'know(Hippos,fly(Rhinos,Aeroplanes))', which cannot be expressed by the existing grammar. But as before, the language can convey 'fly(Rhinos,Aeroplanes)'. However, it can also convey 'know(Hippos,fly(Rhinos,Helicopters))', which is closer in structural size. In this case, which is the closer meaning? Thus finding the closest expressible meaning would require us to formalize the notion of similarity.

We do this by choosing the parse tree with the least number of generalizations needed, starting from the base of the parse tree (outermost predicate). A section of a parse tree attempting to produce a corresponding meaning contributes a distance equal to the minimum sum of distances of its individual branches. Also, if a section of a parse tree cannot even partially match the corresponding submeaning (without any incon-

sistency at any position), it contributes a distance equal to the number of unmatched arguments and predicates in the intended partial meaning. Thus 'know(Hippos,fly(Rhinos,Helicopters))' or 'know (Hippos,Rhinos))' are the closest expressible meanings, with a distance of 3 each since they can match a relation and an argument each ('know' and 'Hippos'), but not'fly(Rhinos,Aeroplanes)'. No other parse trees can match any argument or predicate without an inconsistency at any position. With this notion of meaning similarity, the pseudocode for the invention algorithm is given below. The procedure determines if a given meaning is producible by the agent's grammar, and if not, invents a new signal generated from the closest meaning.

```
PROCEDURE DEFINITION:
proc Invent(Target_meaning, Current_symbol){

boolean=0;          //Can meaning be produced without invention?
ut=RandomString;    //Produced utterance
min= Infinity;      //Distance of closest producible meaning.

for i=1:Number of production rules in Grammar G,
{
C=Category symbol of rule i;
S=Partial semantic meaning on LHS of rule i;
RHS=Partial signal on RHS of rule i;

//For a rule with matching symbol:
if(C ~= Current_symbol) continue; end

utemp=RHS;
Replace variables in 'S' with corresponding category symbols from 'RHS';

[b(1):b(k)]= Atomic elements or category symbols in the semantic S.
// which should not be more than that in Target.
[a(1):a(k)]= Corresponding k elements in the 'Target', determined by paranthesis.
```

```
//some might be composite.

//Match all k tokens:
flag=1; diff=0; //least number of tokens that ‘i th’ rule mismatches
for j=1:k {
if(b(j) is an atomic meaning){
if(a(j) ~= b(j)){
flag=0; diff=numk;
utemp=RandomString; break;}
 }else if (b(j) is a category symbol){
[boolean2 min_i uti]= Invent(a(j),b(j));
if(~boolean2) flag=0; end

diff=diff+min_i;
Replace category symbol b(j) in ’utemp’ by ’uti’;}
}

//Was this rule closer in meaning?
if(diff<min){
min=diff; ut=utemp;}

//On finding working rule, dont search further
if(flag==1){
min=0; boolean=1;
return [boolean ,min, ut];}
}
return [boolean, min, ut];

CALLING THE PROCEDURE:
[bool, dist, utterance]= Invent(Target_meaning, ’S’);
//’S’ is start symbol in the grammar.
```

3.2.3 Interaction

Thus, we now have mechanisms that ensure that any 'speaker' can produce signals for any valid meanings in the meaning space. The 'listener' attempts to correctly parse these meaning-signal pairs. Since at the beginning, the listener's grammar is empty, it can't match the heard utterance meaning pairs initially. However, at the end of each unsuccessful communication, the listener uses the new data [3], to modify its linguistic hypothesis according to the seen data; and thus improves its linguistic proficiency with experience. The speaker too calls the induction algorithm every time the existing linguistic hypothesis is unable to convey a meaning and must call the invention procedure. Thus, the crux of the entire process is the induction mechanism, through which agents can infer from strings the abstract meanings they represent. The following pseudocode summarizes a language game of N interactions between two agents.

```
proc Interact(A, B, N);
%N interactions between two agents A and B, where A is speaker and B is listener.
G1=LoadAgentGrammar(A);
G2=LoadAgentGrammar(B);

for i=1:N{
    Meaning=ChooseRandomMeaning;   //Agent A to choose a meaning
    //Can Agent A convey the meaning? If yes, speak.
    [b1 min_difference speech ]=Invent(Meaning, 'S');
    if(~b1){
        //Agent A needed to invent speech, needs to generalize new example.
        AddNewRule(Meaning-->speech,G1);
        InductionAlgorithm(G1);}

    //Agent B : listener
    b2 = CanIParse(s, speech, G2);
```

[3]The new data tuple (meaning,*signal*) is incorporated in the agent's grammar in the form of a new production rule: New Category Symbol/ meaning $\longrightarrow$ *signal*

```
    if(~b2){
        //Agent B could not parse new speech, needs to generalize new example.
        AddNewRule(Meaning-->speech,G2);
        InductionAlgorithm(G2);}
}
```

At this juncture, we observe that there is an implicit assumption of meaning transference in the model of the language game. In other words, the listener always knows exactly what the speaker's intended meaning was. This is the classical 'gavagi' problem, illustrated by Quine [41]. The context of the illustration is a speaker pointing in the direction of a rabbit, and making the utterance 'gavagi'. The problem is that 'gavagi' could refer to the rabbit, its gait, color or shape; or any other object in the context. Hence, the issue is inferring the meaning of an unknown signal from a set using the context.

3.3 Induction algorithms

The acquisition of I-language given examples of E-language revolves around inductively learning 'good generalizations' which correctly map phonetic strings to semantic meanings. In this section, we explain the induction mechanisms in our model, and how they modify an agent's grammar.

In our formalism, new data is incorporated by a learner as a pairing of a sentence (string of terminals), and the associated semantic structure. However, agents have powers to modify existing rules to subsume new linguistic examples. Rule subsumption in the model operates greedily, in the sense that rules can be modified or created based on observing a single new example at a time, without a provision of a look-ahead or a roll-back mechanism. Essentially, there are three basic operations to generalize rules in a grammar:

3.3.1 Rule merging

Rule merging operates by merging two category symbols as one, if they are seen to have identical usage under any context. For instance, consider the case where there are two rules which are identical except in the occurrence of two category symbols. A simple example is shown below.

$$\text{M/Rhinos} \longrightarrow abc$$
$$\text{N/Rhinos} \longrightarrow abc$$

In this case, the rule merging operation unifies the two categories in the grammar. This is done by replacing all occurrences of one, say N, by the other. Thus, the rule merging operation is a simplistic way of combining word categories, based on overlap in usage.

In a sense, this simple merging model highlights a shortcoming in the generic model, namely that it is ungrounded in perception and the treatment of meanings as discrete, atomic concepts leads to a slightly impoverished treatment. A more realistic scenario would allow merging of category symbols or the categorization of a new meaning to a symbol, based on perceptual similarities between the meanings that they represent, and not just observed usage. For example, if the current grammar has a rule like 'M/Cow $\longrightarrow$ *cow*', and a new observation leads to the incorporation of the following rule 'N/Buffalo $\longrightarrow$ *xyz*' with a new category symbol N, similarity of perceptual features of the meanings 'Cow' and 'Buffalo' can lead to identification of N as category M. In this case, the category symbols would be closer to semantic categories, rather than just linguistic units, as in the current case, and such a model can lead to facilitation of propositional queries based on first order logic.

Additionally, the rule merging operation ensures deletion of duplicate rules.

3.3.2 Rule chunking

Rule chunking operations are provided to enable the agents to generalize. Essentially, the chunking operation considers pairs of rules, and looks for the *least general generalization*

that could be made, and which still subsumes the rules within prespecified constraints. The following two examples illustrate the situation.

Example 1:

S/likes(Mary, Gavin) $\longrightarrow$ *lykesmarygav*

S/likes(Mary, John) $\longrightarrow$ *lykesmaryjon*

In this case, there is a lexical as well as a semantic similarity in the two linguistic examples. Hence, this gets subsumed by replacing the meanings 'Gavin' and 'John' by a semantic variable, and by introducing a new category symbol which can represent 'Gavin' or 'John' by corresponding phonetic subsequences learnt from these examples.

After chunking:

S/likes(Mary, 1) $\longrightarrow$ *lykesmaryN/1*

N/Gavin $\longrightarrow$ *gav*

N/John $\longrightarrow$ *jon*

The original rules are replaced by the above in the agent's grammar. The new set of sentences the agent can parse is now a superset of what it could parse previously.

Example 2:

S/knows(Peter, 1) $\longrightarrow$ *peteknowsA/1*

S/knows(Peter, Heather) $\longrightarrow$ *peteknowsheath*

As before there is a co-occurring lexical and semantic similarity, but the mismatch in this case is between a fully specified terminal and a category symbol. So, the second rule is treated as an instantiation of the first and is removed from the grammar. At the same time, the symbol category A is expanded so as to subsume the other rule.

After chunking:

S/knows(Peter, 1) $\longrightarrow$ *peteknowsA/1*

A/Heather $\longrightarrow$ *heath*

The deliberate mapping of parts of a meaning to parts of phonetic sequences also denotes an explicit bias of the learning algorithm towards compositionality. More significantly for our purpose, the rule chunking mechanism provides agents with the power of generalization.

3.3.3 Generalization

A further important operation is allowed to further assist generalization of rules. Whenever possible, the induction algorithm tries to simplify rules using smaller rules already present in the grammar. For example, assume that the following rules are part of an agent's grammar at some point.

Example:
S/hates(Gavin, John) $\longrightarrow$ *petehatzjon*
A/John $\longrightarrow$ *jon*

In this case, there exists a smaller rule such that its semantic structure is consistent with a part of the semantic structure of the first rule. At the same time, the phonetic sequence corresponding to its meaning occurs as a subsequence in the utterance of the larger rule. The induction algorithm simplifies the larger rule, while the smaller rule remains unchanged.

After generalization
S/hates(Gavin, 1) $\longrightarrow$ *petehatzA/1*
A/John $\longrightarrow$ *jon*

These rule subsumption procedures in the induction algorithm are repeatedly called until the grammar cannot be subsumed further.

3.3.4 Evaluating Kirby's stand

At this point, it would be worthwhile to debate Kirby's claim that compositionality and recursion are natural artefacts due to pressures of a transmission bottleneck, that emerge independent of the learning algorithm. It is indeed true that fewer training examples favour compositional rather than holistic language models and also that linguistic rules and replicators that are more general and show compositionality are more likely to persist. However, in this case at least, compositionality and recursion seem to be heavily favoured by the induction algorithm itself. This is most explicit at two junctures, the first being the rule subsumption mechanism where the least general generalization is used to map parts of the semantic space to parts of the signal space. Such a mechanism, by its very definition, entails compositionality.

Secondly, there is undeniably a strong bias towards compositionality in the invention mechanism used by speakers. In choosing part of the utterance with the 'most similar' meaning as a component in the new signal, the formulation guarantees that the induction algorithm can generalize due to simulataneous overlap of phonetic and semantic structure. While the bias seems natural and quite intuitive, the partisan nature of the induction algorithm means that it can itself be construed as the biologically innate LAD, that Kirby argues against. Indeed, it is difficult to conceive how Kirby's claims of natural emergence of compositionality and recursion might work for a general cognitive model.

Additionally, the formulation makes strong assumptions about a meaning representation that is already in place, without any suggestions about how such a representation might evolve. Also, it is not clear how languages would evolve if the meaning space was differently structured. However, what is undeniable is that the formulation does replicate complex features of natural languages such as compositionality and recursion, even if within a limited framework. Still more importantly, as far as we are concerned, it shows potential for languages adapting and evolving, emergence of competing word orders, and word categories, such as nouns, verbs and fillers, corresponding to semantic categories in the internal representations. In terms of investigating these features for human languages this model is sufficiently powerful, while being tractable since the lin-

guistic hypothesis of any agent at any point can be recorded conveniently as its grammar at that point.

3.4 Extension to a population

3.4.1 Single agent runs

In an Iterated Learning setting with a single agent in each generation, the model described above develops compositional and occasionally recursive languages over many generations. Kirby starts with a meaning space with 5 relational concepts each requiring two distinct arguments, and 5 arguments. The total size of the meaning space is thus $5 \times 5 \times 4 = 100$. Each agent receives only 50 linguistic examples from its parent in its listening phase. Thus a holistic language can never express more than half the possible meanings. The progression of the simulation is marked by an initial phase of mostly holistic rules, and an expanding grammar size. Eventually however, most grammars collapse to about 12 rules, and show compositional structure with clearly identifiable noun and verb classes. The following graph shows progression of several runs in these settings, as documented by Kirby.

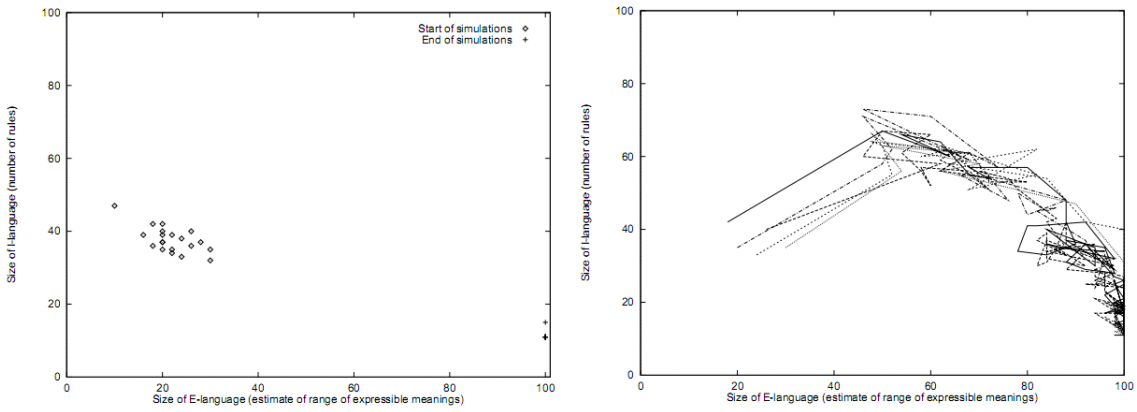


Figure 3.1: Progression of single agent runs in Kirby model (figure from [1])

The compositional languages are marked by a smaller size of I-language (fewer rules), and higher expressivity in the E-language(more conveyable meanings). Most simulation runs reach compositional languages relatively quickly (generally within 20 generations).

```
A\  John     --> gg
A\  Pete     --> cghkbceb
B\  loves    --> fill
B\  knows    --> e
S\  1(2,3)   --> A\2B\1A\3e
A\  Gavin    --> k
A\  Heather  --> c
A\  Mary     --> cgh
B\  doesntknow  --> e
B\  likes    --> adn
B\  hates    --> kb
```

Figure 3.2: Example of an emergent grammar from a single agent population

However, even with a single agent, the stochastic nature of the meaning generation, and the random string generator can lead to different convergence times for different runs [4].

3.4.2 Multiple agents and probabilistic production

The above-mentioned model,as proposed by Kirby, is restricted in several ways. Most significantly, the rudimentary ILM suffers from an impoverished treatment of population dynamics. Realistic populations consist of relatively large communities of agents. Multiplicity of meaning-to-signal mappings within a single agent, as well as and the variety of interactions due to a significant population can lead to significantly different language acquisition compared to learning from a single parent. Secondly, agents in a population dont interact uniformly, i.e. social networks are not homogenous. As surveyed in Section

[4]In our runs, several simulations reached convergence after a couple of hundred generations. In the graph shown, one of the runs (which has a larger grammar size at the end of the simulation than others) reaches a stable grammar after several thousand generations.

1, the structure of social topology can significantly affect the acquisition and evolution of language. Investigating heteregeneous populations can afford intriguing scenarios such as multiple prevalent linguistic hypothesis in a community, and a simultaneous existence of overlapping yet distinct linguistic behaviours. We seek to extend the basic ILM to more realistic population sizes and structures.

3.4.2.1 Population model

In the extended model, let us suppose there are N niches in a social community. Each niche is inhabited by two agents : a parent (belonging to generation i) and a child (belonging to generation $i+1$). However, a listener can receive its linguistic input not just from its parent, but from several speakers in the previous generation. This is dictated by the social topology, which imposes a distribution for communication for each agent. In general, the distribution is not uniform, i.e. each agent is more likely to communicate with some agents than others. Let p_k denote the distribution for the k^{th} agent, where the j^{th} component of the pmf denotes the probability of the agent interacting with the j^{th} agent of the previous generation, when acting as listener.

The simulation still follows an Iterative setting, and there is a turnover of the entire population after each agent has received the specified number $n(50)$ of linguistic examples for induction. The following procedure illustrates the simulation process in a population with N agents in each generation.

```
proc PopulationSimulation(start_gen,end_gen, n, N, P);
//P is a N*N matrix with the i th row equal to p_i.
//n is the number of training interactions for each agent in a generation.

for i=start_gen:end_gen{
    //For each generation:
    parent_gen= i;
    child_gen= i+1;
```

```
    for j=1:N{
       //Each agent acquires linguistic input consisting of n examples.
       for k =1:n/m{
           //m is number of linguistic examples in a basic unit of interaction.
           Sample a number l from the set [1,2,...N] from distribution p_j.

           //Agent l interacts with Agent j
           Interact(l_{i}, j_{i+1}, m);
        }
     }
}
```

In our experiments, we represent different social communities as graphs. A node denotes a social niche, and an edge from one node to another signifies that an agent from one node can listen to or speak to an agent on the other node. All nodes are taken to be self-connected (edge not shown) indicating that a child can always learn from its biological parent. The following simple graph illustrates the situation.

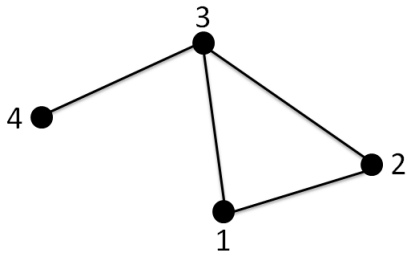


Figure 3.3: Graphical representation of a social community.

Here, a listener at position 2 for example ,receives on an average $\frac{1}{3}^{rd}$ of its linguistic input from interactions with its mature parent at node 2, $\frac{1}{3}^{rd}$ with a speaker at node 1, and the remaining $\frac{1}{3}^{rd}$ from interaction with the mature agent at node 3.

3.4.2.2 Production model

A more subtle nuance that is seemingly hidden in Kirby's simulations is the issue of deterministic production. In Kirby's world, speakers use the same meaning-to-signal mapping to convey meanings. Even if an agent's grammar allows for several ways to produce a meaning, the same utterance is always used in the speaking process. [5] A possible treatment would be to maintain several competing production mechanisms to convey a meaning, and choose the mechanism for production probabilistically from among these. The components of the probability mass function could be functions of some *fitness function* of the production mechanisms, such as the probability of an utterance produced from the mechanism of being understood by a listener. Alternative fitness functions could target simplicity of the signal produced , or the number of rules involved in the production (limiting production cost).

Within the Iterated Learning setting however, all interactions are between fully developed speakers and listeners whose linguistic hypothesis are still nascent. Since listeners always start with blank grammars, no production mechanism can possibly show much success, other than by chance. In such a scenario, weighting a production mechanism on the basis of communicative success doesn't make sense, though it would be the natural choice if the interactions were primarily horizontal. Instead, we weight competing hypothesis on the basis of signal simplicity. As we shall see in the next section, in fact, a bias towards shorter productions is *required* by the model to entertain multiple hypotheses.

3.4.3 Word length and the issue with multiple hypotheses

The induction algorithms described previously allow the length of phonetic utterances to increase as linguistic evolution proceeds. This is because the greedy nature of the induction algorithms can lead to inappropriate generalizations, but in the absence of a roll back mechanism, the generalization can only be corrected through future subsump-

[5] Kirby does this by imposing an ordering on the rules of a grammar, and always using the *first* set of rules the speaking algorithm can find, which can produce the intended meaning.

Parent 1	Parent 2
S/likes(Mary, 1) $\longrightarrow$ *marylyk*A/1	S/likes(Mary, Gavin) $\longrightarrow$ *avn*
A/Pete $\longrightarrow$ *pete*	S/hates(Pete, Gavin) $\longrightarrow$ *petehatzgav*
A/John $\longrightarrow$ *jon*	
A/Heather $\longrightarrow$ *heath*	

tions. This process can potentially lead to monotonic increments in word size, which can disappear only through stochastic non-transmission or word-chunking.

The following examples illustrate the possibility:

Child:

Input examples:

S/likes(Mary, John) $\longrightarrow$ *marylykjon* (from Parent 1)

S/likes(Mary, Gavin) $\longrightarrow$ *avn* (from Parent 2)

Inductions made:

S/likes(Mary, 1) $\longrightarrow$ A/1n

A/John $\longrightarrow$ *marylykjo*

A/Gavin $\longrightarrow$ *av*

Here, the child learns from two parents, one of whom shows compositional production similar to English. However, due to a coincidental match between the single phoneme 'n' in *jon* and *abn*; the child learns a longer signal *marylykjo* for the meaning 'John', than its parents. The situation can be further aggravated if the child now hears 'S /hates(Pete, Gavin) $\longrightarrow$ *petehatzgav*' as an example from Parent 2. Since the child has a rule A/Gavin $\longrightarrow$ *av*, whose semantic and phonetic components are both present in this example, the generalization procedure operates to give the grammar of the child agent the following form:

S/likes(Mary, 1) $\longrightarrow$ A/1n
A/John $\longrightarrow$ *marylykjo*
A/Gavin $\longrightarrow$ *av*
S/hates(Pete, 1) $\longrightarrow$ *petehatzg*A/1

Now, if the child is to produce the meaning 'hates(Pete, John)', the produced utterance is the sesquipedalian *petehatzgmarylykjo*. Subsequently, if the listener for the child has a rule such as 'M/John $\longrightarrow$ *j*', the generalization procedure will again operate to form a production mechanism 'S/hates(Pete, 1) $\longrightarrow$ *petehatzgmarylyk*M/1*o*'; making possible further expansion of sentence length.

Rules once expanded are usually persistent, and can only be truncated occasionally by chunking, or die out stochastically due to non-transmission. It must be noted that the catalyst for the expansion is usually inconsistent input, which is due to the presence of two distinct sources of linguistic input in the case here. The probability of making such inappropriate generalizations would increase with an increase in the number of linguistic hypotheses an agent is exposed to.

On the other hand, if an agent is exposed to consistent linguistic examples, such as from a single source (as in the basic ILM), the nature of the invention algorithm almost certainly precludes an explosion in word length. Interestingly, the scenario of extending the basic ILM to a population and having a probabilistic production mechanism both imply exposure of a listener to multiple linguistic behaviours, and at times inconsistent training examples. Thus, issue of occasional increases in word length will arise especially for larger populations.

In isolation the issue may seem no more than a practical inconvenience, but in context of the simulation it violates practical constraints such as limited memory and processing power in agents, suggesting that these physical constraints need to be specified within the system. Moreover, the notion of unlimited word length is unnatural and non-intuitive, as it goes against the general principle of least effort. Thus, it would seem that it is not only convenient for tractability, but natural to impose a bias against long utterances;

since the least effort principle is innately embedded in all biological systems.

3.4.4 Zipf's law, bias for brevity and memory constraints

As discussed earlier, the induction and production models need realistic constraints such as limited memory and processing speed, as also a bias towards the least effort paradigm. Here we discuss possible mechanisms which can achieve the same.

3.4.4.1 Zipf's law

In the context of word-lengths and word-frequencies, several mathematical formulations have been proposed, as well as empirically corroborated in both natural languages as well as random text. Most of these approaches, more or less, adhere to a general Zipf's like relationship [43, 42].

In our formulation, let us assume that the frequency of usage of a word and its length are functionally related. Let y be the length of all words having frequency x. The Zipfian assumption is that the relative rate of increase of word length is directly proportional to the relative change in word frequency. In other words,

$$\frac{\mathrm{d}y}{y} = -\gamma \frac{\mathrm{d}x}{x} \tag{3.1}$$

$$\Rightarrow y = kx^{-\gamma} \tag{3.2}$$

Hence, x is of the form

$$x = Ay^{-B} \tag{3.3}$$

Hence, the frequency of occurrence of a word is inversely proportinonal to the Bth power of its length. Now, the probability of finding a word is directly proportional to its frequency. Hence, the probability $p(y)$ of finding a word will have a similar form.

$$p(y) \propto y^{-B} \tag{3.4}$$

The value of B has been empirically measured by several studies. The value of B is theoretically a positive constant greater than unity. For most natural languages, the value is between 1 and 2 (According to Strauss et al. [44], the value is around 1.04 for the Russian text of 'Anna Karenina', 1.20 for the original English rendition of 'Portrait of a Lady', and 1.98 for the German 'Hansel and Gretel'. Only for two Slovenian and Indonesian texts in his study is the value larger than 2). [6]

The motivation behind following this derivation is that these Zipfian probabilities can be used as the fitness function to choose signals when several mechanisms can produce the meaning. The probabilities of occurrence of signals of different lengths follows the form shown in Equation 3.4

Hence, the relative probabilities of using different possible signals can be calculated simply by considering all possible signals, and normalizing by summing all probabilities to 1. Using this mechanism, longer signals are penalized and there is a bias towards signal brevity. For example, suppose that the value of B is 2 , and the meaning to be conveyed can be generated by three mechanisms which produce signals of one, two and three phoneme lengths. Thus, the probabilities of these three mechanisms being used are in the ratio $\frac{1}{1^2} : \frac{1}{2^2} : \frac{1}{3^2}$, and the components of the probability mass function can be calculated to be $\frac{36}{49}$, $\frac{9}{49}$ and $\frac{4}{49}$.

In our simulations, we used intermediate values of B, and found that signal growth can only be avoided for values of B greater than 3. When the value is lower, the bias is insufficient to prevent growth in signal length. However, with a sufficient value of B, unnecessary signal-growth, as well as non-compositional islands in the resulting grammars are usually avoided.

[6]The value is in the same generic range,but closer to 1, even for the classical interpretation of Zipf's law as relating word frequency and its rank in a text corpus, which follows a similar derivation. In fact, this is found to be true not just for natural languages, but also random English text.

3.4.4.2 Truncation

Another heuristic to constrain memory usage, and assist in containing word length in case of large populations is to enforce a maximum size of word length that a *terminal meaning* can expand to (In this study, this was kept at four). If the terminal mapping gets excessively complicated, and the size of a terminal word expands to greater than the limit, it is automatically shortened by a deterministic approach which always shortens the name to only its first four phonemes. For example, the rule 'M/Mary $\longrightarrow$ *marianne*' in the grammar gets replaced by 'A/Mary $\longrightarrow$ *mari*'. This can be considered as a policy akin to promoting nick-names to avoid remembering longer mappings. Crucially, the nickname is adopted by deterministic procedure meaning that the issue of global consensus on renaming is avoided through this approach.

However, truncation can frequently result in loss of information, due to which convergence may be delayed for long periods of time. In our simulations, this heuristic failed to work for large population sizes.

3.4.4.3 Shortest production

This is the simplest approach to choosing shorter signals in a population of agents. In this approach, the signal production process simply chooses the smallest possible signal at each step of the production process, effectively choosing the production mechanism which produces the shortest signal. This simple bias adequately avoids increase in word length even with large population sizes. However, this defeats the goal of probabilistic production, since only one signal gets chosen every time to convey a meaning, as in case of the basic ILM. As expected, simulations following this approach normally converge the fastest, due to exposure to fewer linguistic behaviour.

The Zipf's law production mechanism approximately tends to this simplistic mechanism in the limit of B increasing to large values. However, if there are two signals with the shortest length, the Zipf production mechanism is equi-likely to choose both, and thus can never be completely deterministic.

3.5 Social Topologies

We explore the model on several different types of topologies. Namely, the following types of networks are investigated:

1. Fully connected topology
2. Random graphs
3. Loosely connected components
4. Linear topology
5. Ring topology
6. Social graphs

We observe the effect of social structure on language acquisition through both individual and population-level statistics. These parameters include rate of convergence, sizes of grammars (number of production rules), expressivity (fraction of the entire meaning space which can be conveyed by an agent's grammar), global coherence (average communication accuracy between two randomly picked agents in a population) and local coherence, which measures average communicative accuracy between an agent and its immediate neighbours. The meaning and significance of these performance measures is explained in the next chapter.

Chapter 4

Results

In this chapter, we present results of the experiments conducted, and attempt to analyze and interpret them. All simulations are run upto 500 generations, and repeated to verify convergence.

4.1 Performance measures

In our simulations, we use the following terminology for evaluating agents' languages:

- **Grammar size:** This measures the size of the I-language for an agent, and is simply the number of production rules in the grammar. Typically, compositional grammars have a smaller number of rules than non-compositional ones. However, compositional grammars may be larger when an agent holds multiple conflicting hypotheses, for example an agent that has to interact with speakers of two mutually distinct languages.

- **Expressivity:** This value signifies the fraction of the entire meaning space which can be conveyed by an agent's grammar at a give point.

  If this value is low, it suggests that the agent usually has to invent new words to communicate. It also implies that the agent cannot parse signals for most meanings. If this value is high (close to 1), it suggests that the linguistic system

of the agent can always produce a signal for a meaning. However, it does not guarantee that listeners would be able to parse these signals, i.e. expressivity does not imply communicative accuracy.

- **Global coherence:** This value signifies the average communication accuracy between any two randomly picked agents in a community. Thus it is a feature of an entire population, rather than any individual agent. A high value indicates that the language is homogenous throughout the entire community.
- **Average local coherence:** This signifies the average communicative accuracy between an agent and its immediate neighbours. This is a more realistic indicator of communicative accuracy since agents in the model react with socially adjacent agents. The value of global coherence can be low even if local coherence is high due to continuous but changing face of language in the social/geographical fabric.
- **Convergence time:** This is simply the number of generations after which average grammar size and global coherence don't change by more than a threshold.

4.2 Fully connected topology

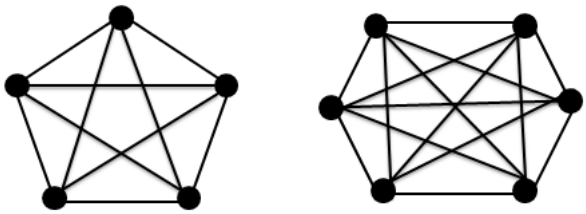


Figure 4.1: Fully connected networks

This is the most basic population model, where all speakers can interact with all listeners. In this case, in a community of N nodes (N agents per generation) a listener, on an average, receives $\frac{1}{N}^{th}$ of its linguistic input from each speaker in the previous generation.

We observe that the our model does , in fact, quite successfully extend the basic iterated learning model to community scenarios. In fact, highly structured, compositional and as well as coherent languages are seen to emerge in the communities.

```
A\  likes    --> l
A\  knows    --> bgh
B\  Mary     --> e
B\  Gavin    --> gfg
B\  Pete     --> c
S\  hates(1,2)  --> gB\1lbB\2b
A\  loves    --> lb
S\  1(2,3)   --> A\1B\2jiB\3
B\  John     --> n
B\  Heather --> e
A\  doesntknow  --> ji
```

Figure 4.2: Example of an emergent grammar from population of size 6

N	Mean Grammar Size	Mean Expressivity	Coherence	Convergence(generations)
1	11.0	1.00	1.00	20.3
2	12.3	1.00	0.92	42.7
3	12.3	0.92	0.87	55.3
4	11.7	1.00	0.91	72.0
5	11.3	1.00	0.81	134.7
6	12.0	0.92	0.77	184.0
7	68.7	0.70	0.24	—

Table 4.1: Results for fully connected networks

We observe the effect of increasing population size on linguistic evolution, and note that for all runs on small population sizes, the grammars that emerge always show almost complete expressivity, compositionality and high values of intra-generational coherence. However, the convergence time increases with the number of nodes in the clique, and the communities fail to reach coherent linguistic systems beyond a certain population size. In the present forumulation this limit is reached with a community size of 6.

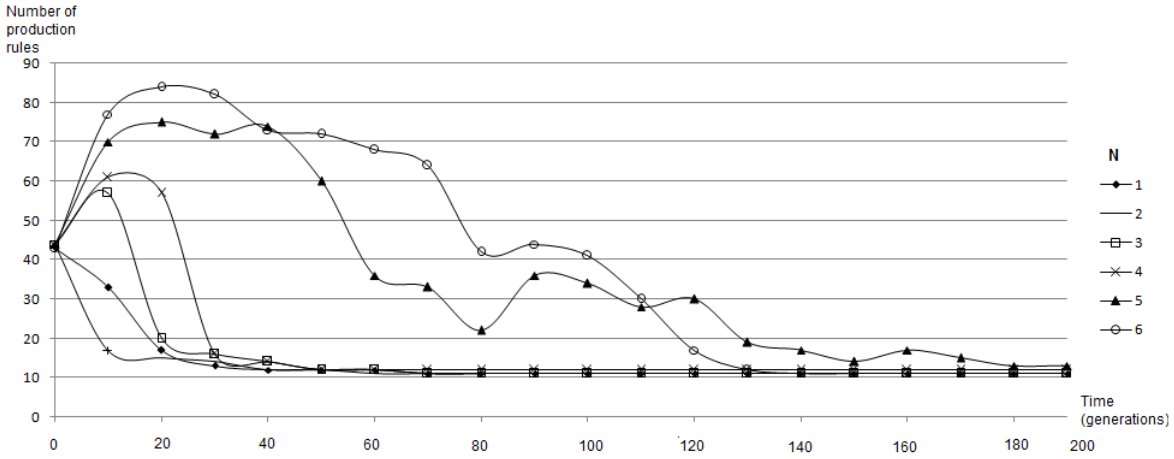


Figure 4.3: Effect of population size on convergence

The failure of large cliques to converge may be because each listener has N sources from where it receives its primary linguistic input (PLI). However, the total amount of information it receives is fixed. If there are too many different sources of linguistic examples, there would be too few examples from any individual linguistic system, and the learner would be hard pressed to find similarities to generalize. In fact, such a learner would never move beyond purely holistic mappings. For a child agent to generalize from multiple existing linguistic systems, there must be adequate representative examples from each. Thus, the result from this experiment is quite intuitive, since it simply states that for effective linguistic acquisition there must not be too many teachers.

4.3 Random graphs

As just seen, our induction algorithm (quite realistically) fails to lead to structured languages if PLI for an agent comes from too many distinct sources. At first sight, this suggests a problem in extending the model to large populations. Real world social topologies, however, do not resemble fully connected cliques. In fact, the number of linguistic sources for any learner is far fewer than N.

We can more accurately simulate real word topologies by replicating our experiments

on random graphs, where the average degree of the graph gives us a handle on the number of linguistic sources that an average agent in the community learns from. We vary this value between 1 and 5, since experiments with a fully connected network already suggested that 6 linguistic sources are too many. For our simulation, we choose a community size of 10 nodes.

We construct the graph by considering all pairs of nodes, and add an edge to each pair selected from a uniform distribution. Note that there is always an edge between a node and itself to start with, indicating that the lineal bond between a parent and its child is always present. We add additional edges till we reach the required value of average degree. Figure 4.4 shows a representative graph used for the simulation.

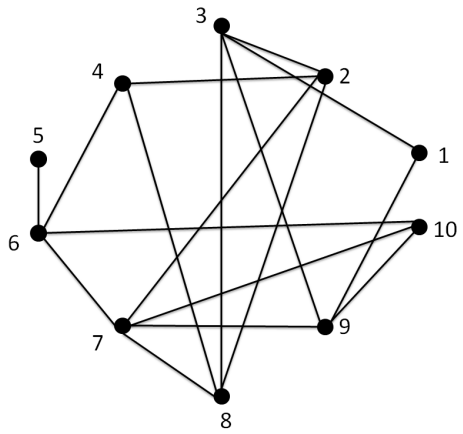


Figure 4.4: Representative sample of a synthetic random graph

Results of the runs for five random graphs of different degrees are summarized in Table 4.2. Most significantly, reducing the number of linguistic sources leads to the emergence of structured and compositional languages in this larger population (10 agents per generation), and possibly suggests a key to scaling the extended model to even larger populations (as we shall see in the section on real world networks).

The crux of the approach is that while the community population maybe large, each agent gets its linguistic input from relatively few sources, such that the data from each

Avg Degree d	Avg Grammar Size	Avg Expressivity	Global Coherence	Local coherence
0.0	11.5	1.00	0.09	1.00
1.6	16.9	1.00	0.65	0.95
2.0	19.1	1.00	0.81	0.91
3.4	24.3	0.97	0.54	0.87
4.4	32.7	1.00	0.29	0.54

Table 4.2: Results for random networks

individual source is adequate to make generalizations. While it could be worthwhile to explore whether increasing exposure in the homogenous case would lead to convergence; in reality, the amount of primary linguistic input can be assumed to be approximately constant, which would need to be shared in case of multiple linguistic parents.

An observation of note here is the difference between the global and average local coherence (This is a departure from homogenous topologies, where the two, by definition, are the same). In this case, while local interactions lead to emergence of significant global coherence; there can be loosely connected features such as small length chains or sub-communities (due to random nature of the social graph) that are relatively isolated (especially for low degree random graphs), and can develop local linguistic traits such as word-orders not prevalent elsewhere in the community. [1] This is a generic feature of non-homogenous networks, since heterogeneity allows for different parts of the community to develop local non-overlapping traits incomprehensible to agents from other parts of the community. For higher degree graphs, the difference between the global and local connectivity is expected to be less significant, as the random graph approaches full connectivity (Figure 4.5).

The runs with different values of the average degree lead to a more interesting observation. In the first case,($d = 0$), the social graph consists of isolated nodes and listeners only learn from their parents at the same node. Figure 4.5 suggests that there is an

[1] In this sense, we look at random graphs as a first approximation to social netweorks. However, random graphs lack characteristics of social networks such as small-worldness and scale-freeness.

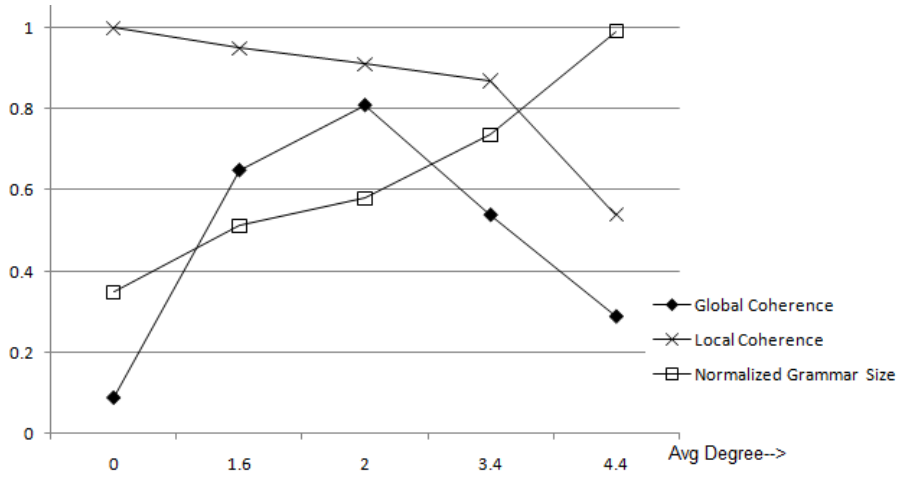


Figure 4.5: Effect of average degree on grammar size and coherence

optimum level of diversity of linguistic sources that an agent should have for a globally coherent language to develop most effectively and quickly. An agent with too many linguistic sources finds it hard to generalize; and takes a long time to converge. On the other hand, with too few teachers, there is too little variety in exposure; the language learnt is primarily lineal; and is not useful for social interaction outside the family, as indicated by the low global coherence. In fact, this situation is akin to each family having a language of its own, and no significant community language as such.

The intermediate sweet-spot indicates a situation where the number of sources is not too high as not to be easily generalizable, while at the same time a learner is exposed to a variety of prevalent linguistic behaviours. In this scenario, a coherent global language evolves in the community.

4.4 Chain topologies

4.4.1 Linear topology

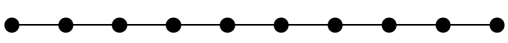


Figure 4.6: Linear topology

This social topology is shown in the Figure 4.6. The community size is 10 (10 agents in each generation), which forms a chain where each node has two neighbours. Thus, a listener receives its linguistic input from three agents, namely its biological parent (resident at the same node) and speakers from the two adjacent nodes. However, listeners at end nodes have only two linguistic sources.

In this case as well, compositional structures emerge , and emergent languages show high communicative potential between neighbours. However, the linguistic hypotheses show a dynamic continuum with respect to topology, and languages at both ends of the chain are widely disparate, with only marginally overlapping vocabulary. The spatial change is exhibited in Figure 4.7, which plots average intra-generational communicative accuracies of Agents at nodes 2 to 10, with the agent at the first node, at the end of one particular simulation run. In all cases, the graph followed a very similar pattern. Results in Table 4.3 are averages for three different runs.

Avg Grammar Size	Avg Expressivity	Global Coherence	Local coherence
22.9	0.94	0.54	0.87

Table 4.3: Results for linear topography

Communicative coherence decreases steadily as social distance increases, and soon dips to near zero-level. However, as seen in the table, local coherence between any two neighbours is high. The situation is not unlike regional dialects of a natural language; where geographically closer regions share more features, while the same language at a distant locale might take an entirely different form. This is true for instance in the case of English in the British Isles, where from North to South the language shows strong Gallic, Scottish and traditional British features in both syntax and diction.

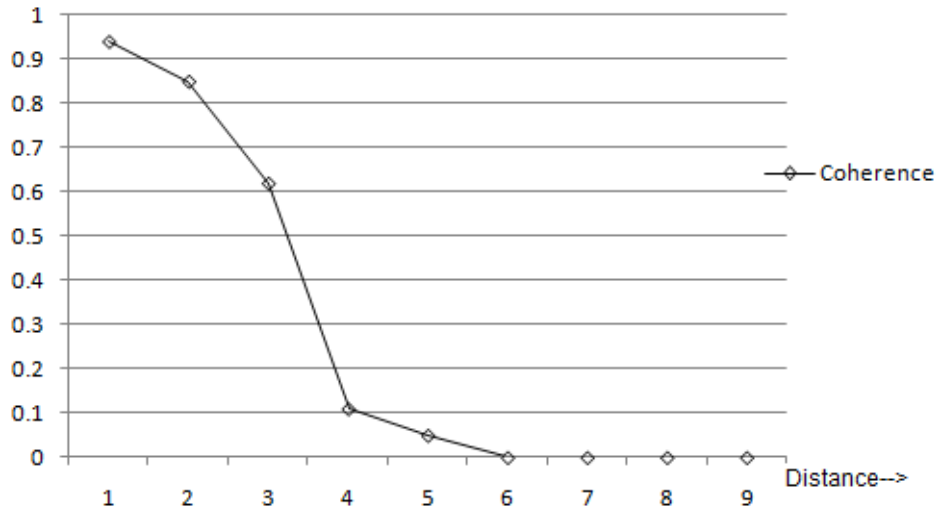


Figure 4.7: Graph of coherence vs social distance in linear topology

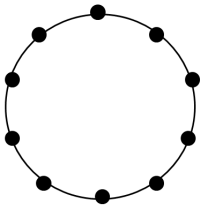


Figure 4.8: Ring topology

4.4.2 Ring topology

The ring is similar to the linear topology, except that the end points are connected with each other. While the language changes as we move spatially, there is smooth metamorphosis underlying high local coherence. Thus any two agents at opposite ends of the ring are almost mutually unintelligible.

Avg Grammar Size	Avg Expressivity	Global Coherence	Local coherence
20.9	1.00	0.67	0.84

Table 4.4: Results for ring topography

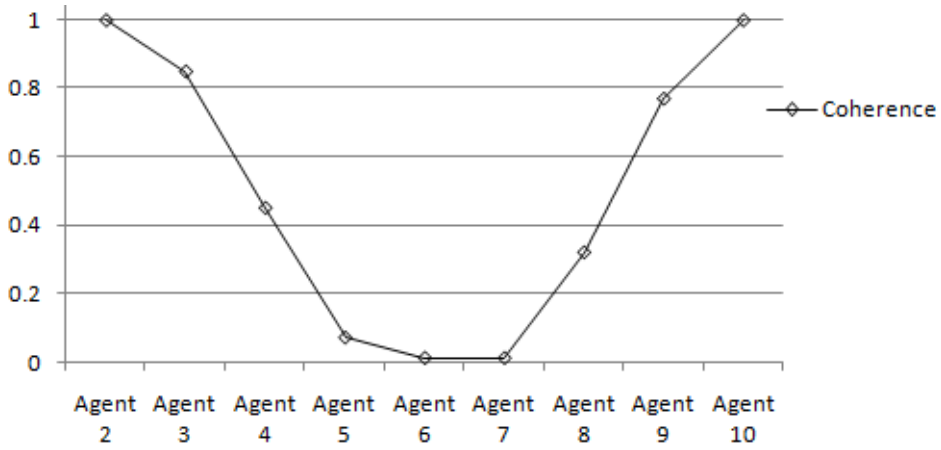


Figure 4.9: Graph of coherence vs social distance for Agent 1

4.5 Weakly connected subgraphs

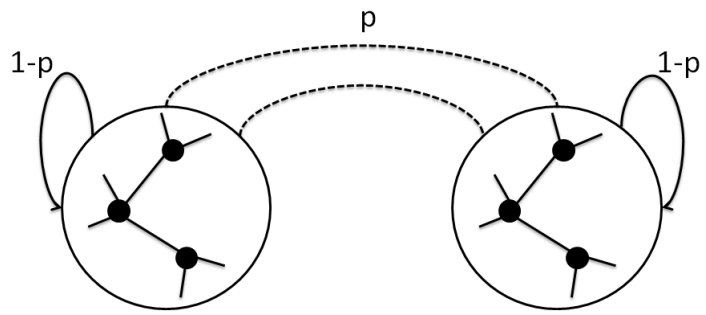


Figure 4.10: Weakly connected communities

The context of this population structure is of a social structure of size N divided into two communities. Agents mostly interact within their own community, but with a small probability p, they occasionally interact with a random member from the external community . In our runs, N equals 10 and each smaller community is a random graph of size 5. In 3 separate runs of this experiment, the value of p is varied between 0.0 and 0.5. Average results of the simulations are summarized in the table below.

p	Coherence in Community 1	Coherence in Community 2	Global coherence
0.1	0.83	0.80	0.41
0.2	0.81	0.84	0.43
0.3	0.84	0.80	0.49
0.4	0.84	0.79	0.76
0.5	0.79	0.85	0.82

Table 4.5: Results for two weakly connected communities

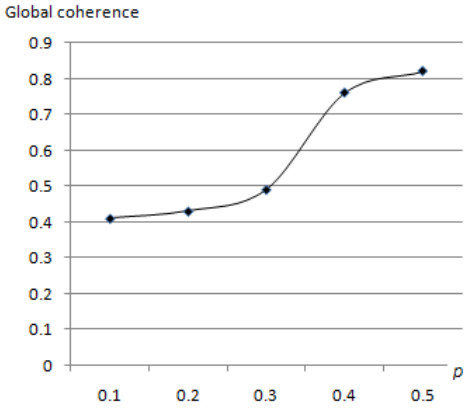


Figure 4.11: Effect of community inter-communication on coherence

Both communities develop strongly compositional and largely coherent languages within communities. Inter community coherence increases with the value of p. Since the two communities are equal in size, if the value of p is close to 0.5, the agents are interacting outside the community exactly as much as they would if there was no distinction between communities, and hence we expect no distinction in language of the two communities. On the other hand, when p is close to 0, there is effectively no interaction between the communities, and we expect the two languages to evolve independently.

However, the interesting observation in this case is that the pattern suggests an intermediate value of p, after which global coherence increases quickly. A study of the grammar-productions of individual agents showed that for values of p greater than this threshold, the dominant word order in both communities always becomes the same.

4.6 Real World networks

While we have attempted to simulate realistic models of population through several topologies, and have in some measure succeeded in extending our model to realistic community sizes, the endeavour would be very much incomplete without testing the formulation on real world social graphs. Real world graphs exhibit several interesting phenomenon such as the *small world* property(small shortest distance between any two nodes), an adherence to a Power law in degree distribution, *scale-freeness* (exhibiting similar structure at all resolutions), high clustering coefficients, and the presence of community structure; that are hard to replicate in synthetic graphs.

Moreover, these networks are typically considerably larger than the communities we have experimented with till now. In this section, we present results of our simulations on two well known real world social networks. [2]

4.6.1 Krackhardt Office CSS

This network, where edge relations denote friendship, was compiled by David Krackhardt. The data consists of social structure data, collected from 21 management personnel in a high-tech, machine manufacturing firm. The data was collected to study the effect of an intervention strategy by the firm management. Each person in the survey was asked for every person X in the office:'Who is a friend of X?' Thus, every person revealed his or her friendship relationships, and perceived friendship relationships between others.

For our study, we take the input given by each person, which consists of a 21×21

[2]The power law distribution of degrees implies that some nodes in large graphs have very high degrees (> 15). Agents at these nodes would have difficulty in making generalizations, because of too many linguistic sources. For these nodes, the language game is modified to constrain linguistic input from 5 randomly selected neighbours (since 6 was the maximum feasible number of linguistic sources for a listener in a fully connected topology). However, the chosen neighbours can differ for every new listener to ensure that the structure of the topology is respected.

Additionally, we have used graph datasets where the structure is not extreme in this sense, and high-degree nodes are relatively few.

matrix A, where a non-zero entry for A_{ij} signifies that person j was perceived to be a friend of person i . We take a majority decision on the friendship relations. An edge is considered to be present if a majority of the 21 participants (11 persons) perceive the friendship relation to exist. Also, self edges are added for every node, as before.

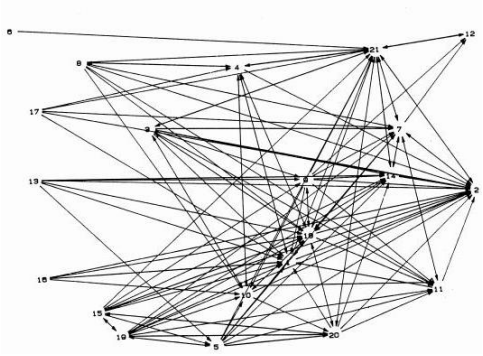


Figure 4.12: Krackhardt's Office Cognitive Social Structure (figure from [46])

Next we generalize the given data such that if X is a friend of Y, then Y is a friend of X. This yields a symmetric adjacency matrix, with more than a hundred edge relations. The consensus structure so obtained (shown in Figure 4.12) is used for the language simulation.

4.6.2 Zachary Karate network

The Zachary Karate club data is a famous social network showing friendship relations between 34 members of a university Karate club. Wayne Zachary (1977) used the data and a conflict resolution model to explain polarization in the group following a dispute. The data is already in form of an undirected graph, with 78 edge relations between the nodes.

4.6.3 Results

Three simulation runs were made on both networks upto 1000 generations. In both cases, some general trends were noteworthy. Most significantly, the large majority of agents developed small stable compositional grammars (around 15 production rules). In spite

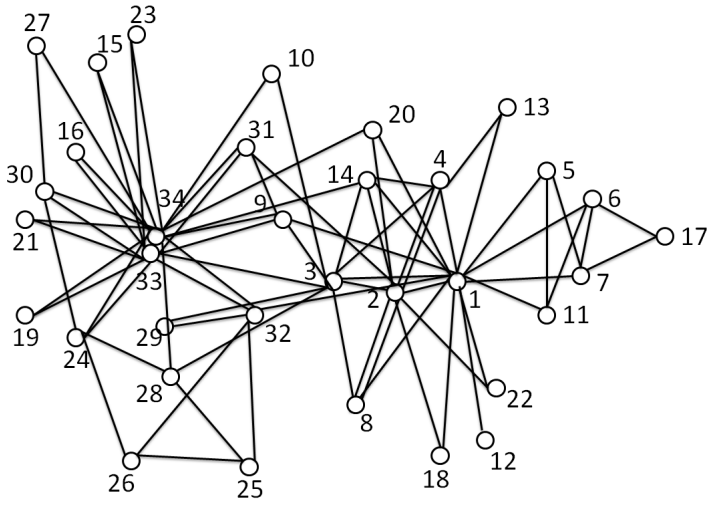


Figure 4.13: Topology of Zachary Karate club network

of the large populations, the networks developed relatively effective languages, as seen by the high average value of local coherence.

However, if agents were allowed to interact with all neighbours, agents with larger number of sources (hubs, such as Node 33 of the Zachary network) showed very large grammars (>40 rules of production), including a few holistic mappings. These nodes also exhibit relatively lower values of local coherence, as expected due to excessive linguistic variety at these nodes.

Avg Grammar Size	Avg Expressivity	Global Coherence	Local coherence
19.8	0.96	0.63	0.82

Table 4.6: Results for the Krackhardt Cognitive Social Structure

Avg Grammar Size	Avg Expressivity	Global Coherence	Local coherence
16.4	1.00	0.55	0.86

Table 4.7: Results for Zachary Karate Network

A closer investigation of the individual grammars reveals several interesting properties. Table 4.8 gives an individual analysis of agents' grammars in the Zachary Karate

network for one particular interesting run. While the table shows that all nodes eventually reach relatively compositional languages, it also suggests some non-trivial observations.

While the SOV and SVO word orders become prevalent ubiquitously throughout the community, the appearanace of OVS and VSO type productions shows a clustering within the community. This is apparent in Figure 4.14 showing the spatial emergence of these two word orders in the community graph. There is a clearly an OVS speaking community to the east of the graph, whereas there is also a community exhibiting the VSO order located more to the west. The fringes are largely untouched by these linguistic features.

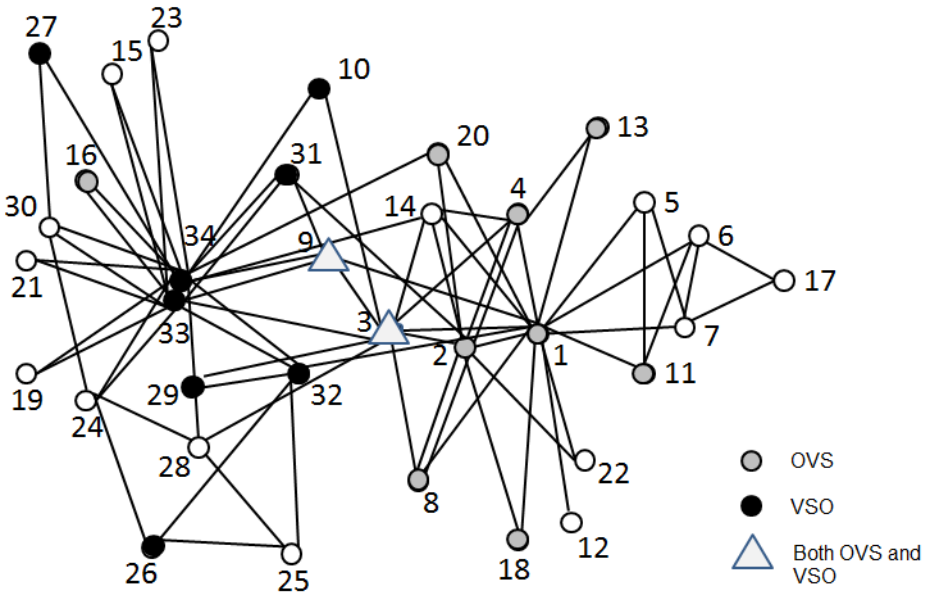


Figure 4.14: Emergence of two word orders in Zachary Karate club network

We also note that there is a partial overlap in the two communities, and Node 3 and Node 9 show developed compositional production rules with both OVS and VSO word forms. Consequently, these nodes also have slightly larger grammars than their immediate neighbours.

Agent	Grammar Size	Expressivity	Preferred word order
1	14	1.00	SOV,SVO,OVS
2	13	1.00	SVO,OVS
3	17	1.00	SVO,SOV,OVS,VSO
4	16	1.00	SVO,SOV,OVS
5	13	1.00	SVO,SOV
6	13	1.00	SVO,SOV
7	13	1.00	SVO,SOV
8	14	1.00	SVO,OVS
9	19	1.00	SVO,SOV,OVS,VSO
10	17	1.00	SVO,VSO,SOV
11	12	1.00	SVO,SOV,OVS
12	11	1.00	SVO
13	12	1.00	SVO,SOV,OVS
14	14	1.00	SVO,SOV,VSO
15	23	1.00	SVO,SOV,OSV
16	19	1.00	SOV,SVO,OVS
17	11	1.00	SOV
18	12	1.00	SVO,SOV,OVS
19	15	1.00	SVO,SOV,VOS
20	15	1.00	SOV,OVS,SVO
21	18	1.00	SOV,SVO
22	11	1.00	SVO
23	16	1.00	SVO,SOV,OSV
24	25	1.00	SOV,SVO,OSV
25	19	1.00	SVO,VSO,SOV

In fact, these nodes represent overlap between two natural communities in the Zachary

Agent	Grammar Size	Expressivity	Preferred word order
26	20	1.00	SVO,SOV,VSO
27	19	1.00	SVO,VSO,SOV
28	21	1.00	SVO,SOV
29	15	1.00	SVO,VSO,SOV
30	22	1.00	SOV,SVO
31	18	1.00	SOV,SVO,VSO,OSV
32	23	1.00	SOV,SVO,OSV,VSO
33	28	1.00	SVO,VSO,SOV,OSV
34	27	1.00	SVO,SOV,OSV,VSO

Table 4.8: Individual agent results for Zachary Karate Network

network. As discussed earlier, the Zachary network was documented with a polarization of agents into two communities supporting the manager and the coach of the club respectively. This is shown in Figure 4.15. Since this is a rare case where ground-truth is conclusively known, the Zachary network has been widely studied in works on overlapping communities. As is clear from Figures 4.14 and 4.15, the particular run remarkably shows almost the same cluster boundaries as the ground truth on polarization. [3]

In fact, a seminal community overlap algorithm by Wang et al identifies nodes 3, 9 and 10 as the overlap between communities. Our set of nodes showing both prevalent word orders is contained within this. However, the social network used in the simulation is not identical to Zachary's graph, since we add self edges to all nodes to ensure lineal descendence of language traits.

The basic observation still is that a community to the right follows OVS type word productions, while VSO type productions are restricted within a community of the left. The identified set of nodes forms the interface between the two communities, serving to transfer linguistic behaviour from one community to the other, and at the same time

[3]While there was some level of polarization in two of three simulation runs on the Zachary network, in the other case the effect was not as distinct and remarkable.

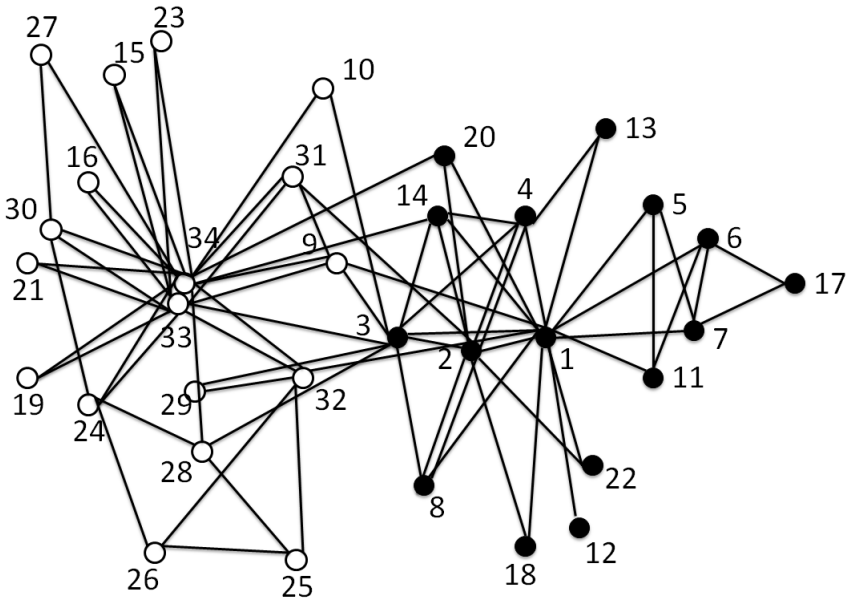


Figure 4.15: Ground truth for Zachary Karate club network

needing to comprehend both word orders. These boundary-nodes hence seem to be linguistically crucial, since they represent a transfer from one language form to another.

Still more significantly, the observation underscores the role that physical topology plays in language evolution. In this case, a non-trivial partition in the physical structure of the graph acts as a seed to bifurcate linguistic behaviour, and eventually leads to the development of two distinct language forms, rather than a single one.

In conclusion, the experiment suggests that the induction model is generalizable to realistic social graphs with a large number of agents, provided the degree of most vertices is not high . The suggested corrective heuristic in the language game is seen to avoid the problem due to the power law property, and agents at nodes with high degrees show similar levels of compositionality as other nodes. Secondly, the structure of the social graph that determines relative interactions of agents can be crucial in creating linguistic divergence.

Chapter 5

Conclusion and Future work

5.1 Conclusion

In this thesis, we have extended an inductive learning model proposed by Kirby to realistic populations and social structures. We have developed a framework where multiple agents can interact in an iterated learning setting, and each agent can receive its primary linguistic input from a set of speakers according to a distribution specified by the existing social topology. We also try to extend the deterministic production model to a probabilistic one. In doing so, we have identified how scenarios maintaining multiple distinct hypotheses can lead to the problem of increasing word size due to the greedy nature of the induction algorithm, and have identified the need to impose physical constraints on memory and processing power. We have suggested how a bias towards brevity can be used as a fitness function for the probabilistic production model, using a Zipfian approach.

An investigation of the extended model on different social topologies led to several insights. It was observed that coherent language develops most effectively in communities that are well connected, but when learners do not get their primary linguistic data from too many sources. This observation led to an extension of the approach to large population sizes. While most simulations on language have focused on the evolution of a coherent lexicon, the current model shows promise in incorporating several aspects such

as syntax and context-dependency at a population level. In chain-like topologies where average shortest distance between two randomly picked nodes is large, the model showed continuous change through the social fabric, while exhibiting local coherence. The model was also extended to real world social networks, leading to interesting observations, and vindicating the significance of social topology in guiding the course of language development.

However, the greedy generalization built into the induction algorithm leads to inevitable problems in scenarios of probabilistic production, unless restricted with a heavy bias. A heavy bias, on the other hand, partly defeats the purpose of having a probabilistic production in first place. Also, while compositional grammars develop in population scenarios, grammars are not as concise for all agents, as in the case of a single agent population. Occasionally, an agent's grammar does not converge (reduce further in size) after becoming fairly compositional. Also, the induction mechanism frequently leads to production rules developing non-compositional islands, especially in intermediate generations.

5.2 Future work

The suggested framework shows promising directions to investigate the development of simple syntax in communities. In simulations in this study, social networks and the interactions they led to were symmetric. The possibility of asymmetric interactions needs to be be investigated. Such a situation may happen due to differences in agents such as social rank, status, age and gender. Within the same framework, it would be worthwhile to investigate the effect of a ubiquitous speaker, which can act as a normalizing agent, similar to the TV or media. Lastly, while we considered scenarios of co-development in two communities, scenarios of evolution of language in merging of two communities can also be studied in the framework.

The framework can also be extended from being purely iterative to a more generic model by adding intermediate stages to the life of an agent. In this way, horizontal

interactions can be accommodated in an incremental manner.

The induction mechanism may also be modified to accommodate look-ahead and roll-back mechanisms, and avoid problems of accidental incorrect generalizations. Alternatively, improvements in the induction mechanism could involve moving from the current greedy mechanism of hypothesis selection (as opposed to the MDL mechanism proposed by Brighton), which is in the flavour of genetic algorithms, towards more concrete mathematical formulations based on optimization approaches.

Finally, to reduce the length of simulation runs we only use meanings with a single relation and two arguments, although the framework allows for more complex meanings. It is worth knowing whether the same behaviour is seen with larger meanings and larger meaning spaces.

Bibliography

[1] Kirby, S. Learning, bottlenecks and the evolution of recursive syntax. Linguistic Evolution through Language Acquisition: Formal and Computational Model, 173203, Cambridge University Press, Cambridge (2002).

[2] Brighton, H., Smith, K., & Kirby, S. Language as an evolutionary system. Physics of Life Reviews, 2:177226 (2005).

[3] Frisch, Karl von. The Dance Language and Orientation of Bees. Cambridge, Mass.: The Belknap Press of Harvard University Press (1967).

[4] Saffran, J., Hauser, M. D., Siebel, R., Kapfhamer, J., Tsao, F., & Cushman, F. Grammatical pattern learning by human infants and cotton-top tamarind monkeys. Cognition, 107:479-500 (2008).

[5] Savage, S., Rumbaugh, D. M., & Boysen, S. Do Apes Use Language? American Scientist, 68:49-61 (1980).

[6] Nollman, J. Who Communicates. The Interspecies News-letter, Winter (1995).

[7] Chomsky, N. Aspects of the theory of syntax. Cambridge, MA: MIT Press (1965).

[8] Osherton, D., Stob, M. & Weinstein, S. Systems that learn. Cambridge, MA: MIT Press (1986)

[9] Uriagereka J. Rhyme and reason: An introduction to minimalist syntax. Cambridge, MA: MIT Press (1998)

[10] Brighton, H. Compositional Syntax from Cultural Transmission. Artificial Life, 8(1):25-54 (2002).

[11] Gold, E. M. Language Identification in the Limit. Information and Control, 10:447-474 (1967).

[12] Vapnik, V. N. Statistical Learning Theory. Wiley, New York (1998).

[13] Steels, L. Language as a Complex Adaptive System. In Schoenauer, M., editor, Proceedings of PPSN VI. Berlin, Germany: Springer-Verlag (2000).

[14] Briscoe, E. J. Language as a Complex Adaptive System: Coevolution of Language and of the Language Acquisition Device. In H. van Halteren and et al., editors, Proceedings of Eighth Computational Linguistics in the Netherlands Conference (1998).

[15] Briscoe, E. J. The Acquisition of Grammar in an Evolving Population of Language Agents. Electronic Transactions on Artificial Intelligence, 3 (1999).

[16] Dunbar, R. Grooming, Gossip, and the Evolution of Language. Cambridge, MA: Harvard University Press (1996).

[17] Nettle, D. Using Social Impact Theory to simulate language change. Lingua, 108(2-3):95-117 (1999).

[18] Ke, J., Gong, T., & Wang, W. S-Y. Language change and social networks. Communications in Computational Physics, 3(4):935-949 (2008).

[19] Ke, J., Minett, J. W., Au, C-P., & Wang, W. S-Y. Self-organization and selection in the emergence of vocabulary. Complexity, 7(3):41-54 (2002).

[20] Smith, K., Kirby, S., & Brighton, H. Iterated learning: a framework for the emergence of language. Artificial Life, 9:371-386 (2003).

[21] Kirby, S. The evolution of meaning-space structure through iterated learning. In Second International Symposium on the Emergence and Evolution of Linguistic Communication (2005).

[22] Pinker, S. The Language Instinct: How the Mind Creates Language. New York: HarperCollins (1994).

[23] Lee, Y., Collier, T. C., Taylor, C. E., & Stabler, E. P. The role of population structure in language evolution. In Proceedings of the 10th International Symposium on Artificial Life and Robotics (2005).

[24] Mufwene, S. S. Competition and selection in language evolution. Selection, 3(1):45-56 (2002).

[25] Mague, J. P. On the importance of population structure in computational models of language change. In Proceedings of the 31st Penn Linguistics Colloqium (2007).

[26] Senghas, A. & Coppola, M. Children creating language: how Nicaraguan sign language acquired a spatial grammar. Psychological Science, 12:323-328 (2001).

[27] Nowak, M. A., Krakauer, D., & Dress, A. An error limit for the evolution of language. Proceedings of The Royal Society of London. Series B, Biological Sciences, 266(1433):2131-2136 (1999).

[28] Komarova, N. L. and Nowak, M. A. Language dynamics in finite populations. Journal of Theoretical Biology, (3):445-457 (2003) .

[29] Niyogi, P. Phase Transitions in Language Evolution. In L. Jenkins, editor, Variation and Universals in Biolinguistics. Elsevier Press (2004).

[30] Batali, J. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In Ted Briscoe, editor, Linguistic Evolution through Language Acquisition: Formal and Computational Models. Cambridge University Press (2002).

[31] Livingstone, D. The Evolution of Dialect Diversity, in Simulating the Evolution of Language. Editors: A. Cangelosi and D. Parisi, London: Springer-Verlag, 99117 (2001).

[32] Steels, L., & Kaplan, F. Bootstrapping grounded word semantics. Linguistic evolution through language acquisition. page53 (2002).

[33] Steels, L. Constructing and sharing perceptual distinctions. In: van Someren, M. and G. Widmer (editors). Proceeding of European Conference on Machine Learning. Springer-Verlag, Berlin. (1997).

[34] Brighton, H. Linguistic Evolution and Induction by Minimum Description Length. In Werning, M. and Machery, E., editors, The Compositionality of Concepts and Meanings: Applications to Linguistics, Psychology and Neuroscience. Frankfurt: Ontos Verlag (2005).

[35] Smith, K., Brighton, H., & Kirby, S. Complex systems in language evolution: the cultural emergence of compositional structure. Advances in Complex Systems, 6(4):537-558 (2003).

[36] Smith, A. D. M. Establishing Communication Systems without Explicit Meaning Transmission. In J. Kelemen and P. Sosk, editors, ECAL01, pages 381-390. Prague: Springer (2001).

[37] Conway, C. M., & M. H. Christiansen. Sequential learning in non-human primates. Trends Cogn. Sci. 5:539546 (2001).

[38] Christiansen, M. H., & J. T. Devlin. Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In: Proc. 19th Annual Conf. of the Cognitive Science Society, pp. 113118. Mahwah, NJ: Lawrence Erlbaum (1997).

[39] Chater, N., F. Reali, & M. H. Christiansen. Restrictions on biological adaptation in language evolution. PNAS 106:10151020 (2009).

[40] Lupyan, G., and M. H. Christiansen. Case, word order, and language learnability: Insights from connectionist modeling. In: Proc. 24th Annual Conf. of the Cognitive Science Society, 596601. Mahwah, NJ: Lawrence Erlbaum (2002).

[41] Quine, W. Word and Object, Cambridge, MA: MIT Press (1960).

[42] Zornig, P., Kohler, R., & Brinkmoller, R. Differential equation models for the oscillation of the word length as a function of the frequency. In: Glottometrika12. Bochum, 2540 (1990) .

[43] Herdan, G. The advanced theory of language as choice and chance. Berlin (1966).

[44] Strauss, U., Grzybek, P., & Altmann, G., Word length and word frequency. In: Grzybek, P. (editor) Contributions to the Science of Language. Word Length Studies and Related Issues, 255-272. Boston: Kluwer (2005).

[45] Kirby, S. and Hurford, J. The emergence of linguistic structure: An overview of the iterated learning model. In Cangelosi, A. and Parisi, D., editors, Simulating the Evolution of Language, chapter 6, 121-148. Springer Verlag, London (2002).

[46] Krackhardt D. Cognitive social structures. Social Networks, 9, 104-134 (1987).

[47] Xiaohua Wang, Licheng Jiao, Jianshe Wu, Adjusting from disjoint to overlapping community detection of complex networks, Physica A: Statistical Mechanics and its Applications, Volume 388, Issue 24. (2009) .

[48] Zachary, W. W. J. Anthropol. Res. 33, 452473 (1977).

[49] Zuidema, W. How the poverty of the stimulus solves the poverty of the stimulus. In: Advances in Neural Information Processing Systems 15, ed. S. Becker, S. Thrun, and K. Obermayer. Cambridge, MA: MIT Press (2003).