Simplicity as a Driving Force for Language Evolution

Ram Ji Tripathi
Y1284, IIT Kanpur

Abstract
Human language evolution is the most discussed issue at present and various theories and models are proposed to study this problem. This paper presents Brighton's theory –Simplicity as a driving force for language evolution in the light of Iterative learning model. Cultural transmission generation to generation is a basic phenomenon which affects language evolution and language may evolve or destroyed due to this complex process. This paper also takes an overview of hallmarks of language and innateness hypothesis and cultural transmission as tool for language evolution.

Introduction
Human Language is unique to humans only; because no other species have language even close human languages. Human language has some distinct and unique features like Compositionality, creativity Cultural transmission, Duality, Semantic Structure, Arbitrariness, open-endedness, finite infinities etc. We compare human languages with its biologically near relatives and find that there is only 1% difference in genomic structure yet languages are poles apart. No animal has a language with compositionality except dance of bees. So it is very interesting to study that how these properties are evolved and why they are unique to humans.

Languages evolved themselves in due course of time. The way by which they evolve themselves is termed as "Language Evolution". Every human language has compositional structure. In terms of compositional properties we learn semantic complexity of language and induction based on sparse language exposure. Based on the observation of universal features of language and to make our study simpler we can propose a universal grammar (UG). UG is an important concept because it takes care of all languages which are spoken at present or lost due course of time. It also takes care of languages which may evolve in future. So in brief it takes care of all possible languages and UG can be taken as object to be explained by the cognitive sciences.

Innateness Hypothesis
Seeing the complexity of human languages, we can assume that it is innate. As Chomsky observed that it is impossible to avoid the conclusion that at least some part of the human language is inbuilt in biological structure. Children learn language in a very fast and efficient way. They learn basics of language in their very early days. It is impossible to learn unique features of language without biological organs as environment can't be source of such fast learning. This argument is known as "Poverty of Stimulus,"

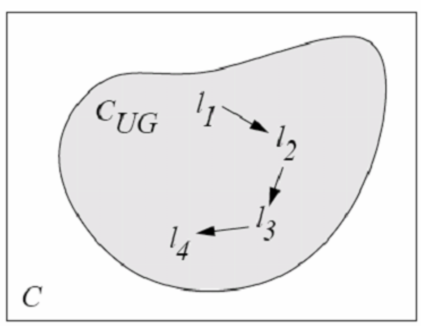
But languages are so much different that everybody can appreciate the question—if language is innate then why they differ so much. To answer this basic question scientists propose that knowledge of language is not encoded in genes instead basic properties of language are encoded in genes. Still it is not clear up to which extent these properties are encoded in genes.

Principle of Detachment

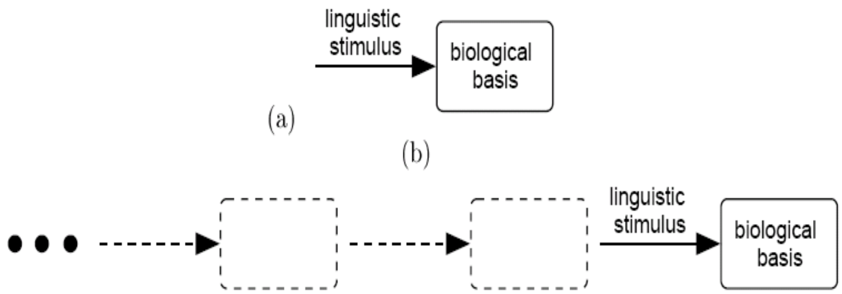
It is clear that we can't explain cognitive processes without innateness argument. So we can safely assume that knowledge of language is the result of biological organs as well learning. This assumption is known as the principle of detachment and it is proposed by Brighton. This principle suggests that biological machinery and learning process both must be taken into account to study language evolution.

Iterative Learning Model (ILM)

Human languages evolve during cultural transmission and adapt themselves. It seems that relevant information for language evolution is encoded within them. To test theories based on these assumption, we have a framework which is known as Iterative Learning Model (ILM). ILM is a generation based model i.e. production of one generation works as input for the new generation and then next generation performs and again outputs its. In this process we see how language is evolved. We don't study behavior of individual agents instead we go for study of whole language from one generation to another. As this cultural transmission from one generation to another generation is not 100% accurate and reliable, Language changes. Some of the linguistic structures survive from one generation to another while others disappear. Hence language changes and slowly evolves. This process is called cultural adaptation. Phenomena of learning few structures of language from one generation to another and leaving few during such transmission is termed as Cultural Selection for Learnability.



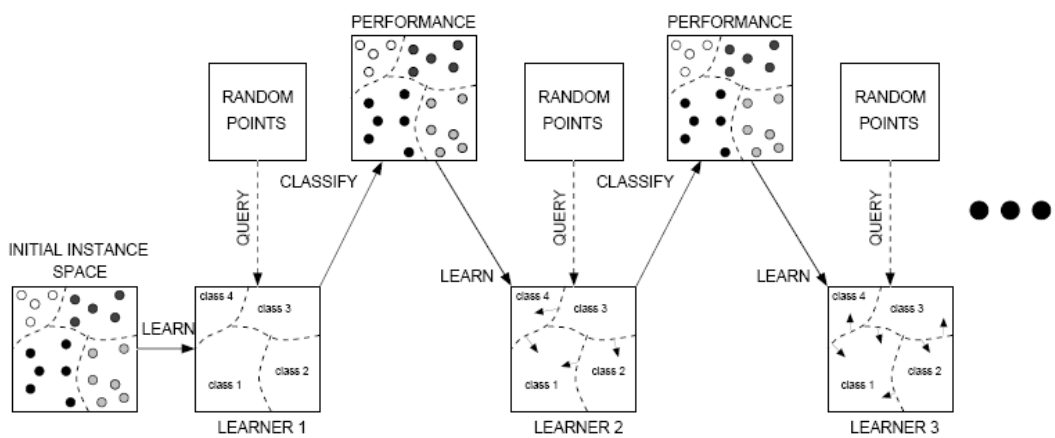
Language changes I1→I2→I3→I4 due to cultural transmission. Here important observation is that all four languages I1, I2, I3 and I4 are completely within the set of UG which justifies the assumption of a UG.



a) Shows a agent of one generation in ILM model
b) Shows ILM model and passing of knowledge of language from generation to another

I-Language and E-Language

Iterative learning model takes biological basis and learning both as important factor for knowledge of language. Hence it assumes two forms of knowledge of language namely internal language (I-language) and external language (E-language). Internal configuration of cognitive structures are taken as I-language while E-language is the externalized linguistic performance derived from internal linguistic competence (I-Language) In terms of I- language and E-language transfer of language from one generation to another is just change of E-language to I-language and then change of I-language into E-language and this cycle goes on from generation to generation.



Picture shows transfer of knowledge of language from one generation to another.

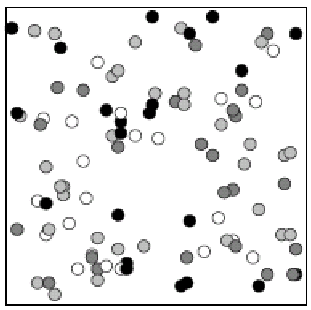
Just understanding of cognitive basis of language is not sufficient to understand evolution of language. We must also learn path of cultural transmission of language and figure out how structure of language is changed. Just seeing the initial language and final language after a number of generations is not sufficient enough to study the evolution process of language. We must have an account of change in language from one generation to another generation

Batali model and Kirby model: a comparative study

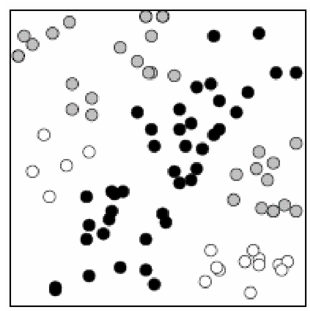
There are two famous models of iterative learning namely Batali Model and Kirby Model. Both models study emergence of compositionality property in language. Batali took some very weak assumptions. He assumed that population is fixed without any turnover. Kirby's assumptions are realistic. He assumes a collection of

agents with turnover. Also in this model after a number of rounds of communication, agents can be chosen randomly and then they can be replaced with new blank agents. Batali used recurrent neural networks to map signals to mapping.
Batali successfully showed emergence of structured signals by interactions of agents. Kirby modeled real life cultural transmission and used the fact that E-language of one generation changes into I-language of next generation and then again I-language is converted into E-language by performance of that generation
Like Batali, Kirby also showed emergence of compositionality, starting from a holistic system. In brief, results of batali and Kirby are same, still Kirby model is more realistic because it takes care of real life scenarios and captures the spirit of cultural transmission.

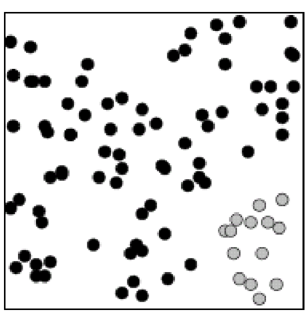
If we take our meaning space as (x, y) in real two dimensional plane and try to evolve signal space by using ILM, we find that results are very enthusiastic as depicted by following figures--



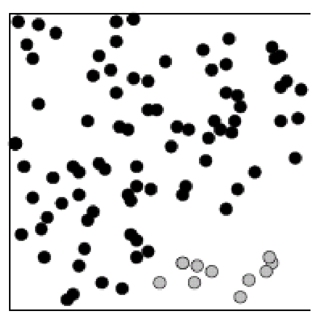
1 ITERATION



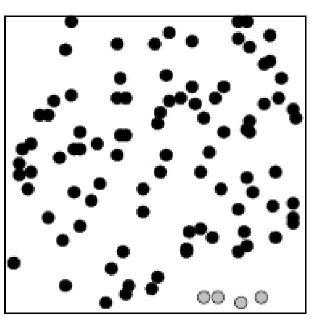
10 ITERATIONS



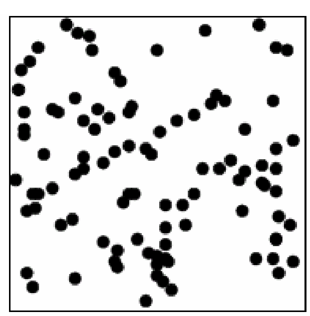
30 ITERATIONS



50 ITERATIONS



59 ITERATIONS



60 ITERATIONS

Here black dots represent those points which are clearly understood by one particular generation i.e. they are matched with their signals and this mapping is understood by whole generation. White dots shows meanings understood by few members of a generation.

a) Meaning space after one iteration.
b) Meaning space after ten iterations.
c) Meaning space after thirty iterations.
d) Meaning space after fifty iterations.
e) Meaning space after fifty nine iterations.
f) Meaning space after sixty iterations.

Language: As mapping of meaning space to signal space

For our purpose to study language evolution, we would like to define language as infinitely large structured mapping from meaning space to signal space. This view of language helps us to study a very interesting property of languages i.e. compositionality. Compositionality means that we can understand meaning of a signal just by understanding its constituents. It's one of the most important properties of human language. In compositional language, similar meanings will always map to similar signals. This property is known as neighborhood preserving property. It is based on simple observation that as similar meanings have some common constituent which will map to same signal hence similar signals must be in neighborhood as meanings are.

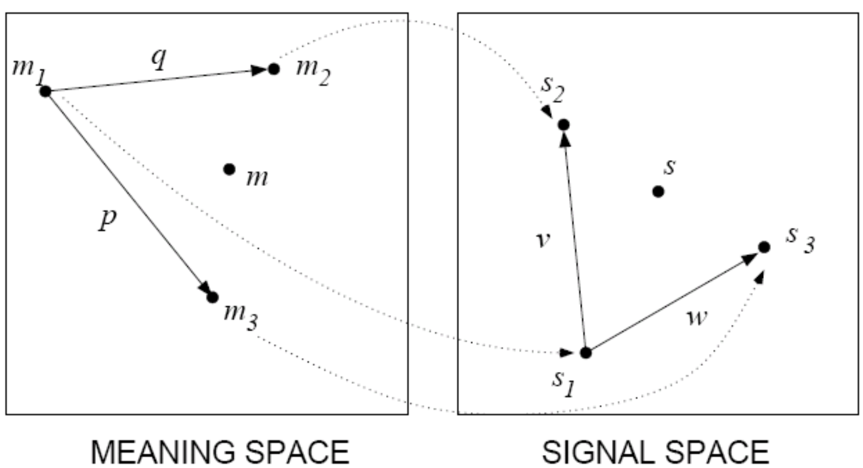


Figure depicts language as mapping between meaning space and signal space. It also shows neighborhood principle by mapping 'm' to's'. 'm' is close to 'm1', 'm2', 'm3' in meaning space hence 's' is close to 's1', 's2 ', 's3' in signal space.

Simplicity

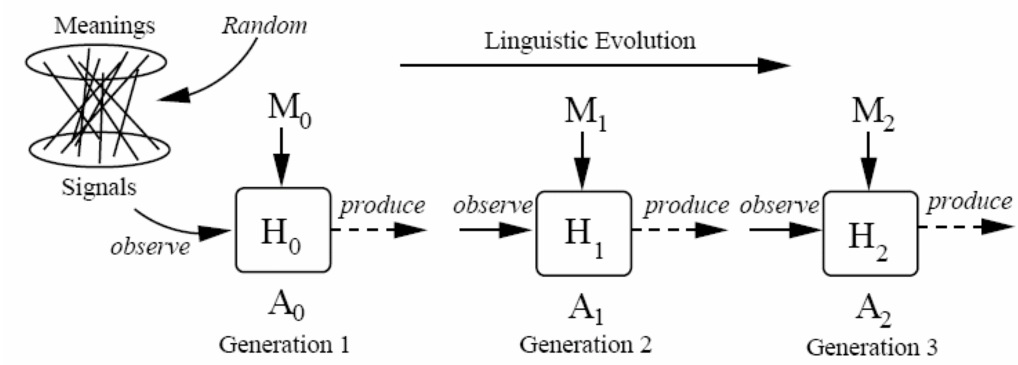
Simplicity here means minimal in some sense. It is of two kind's namely explanatory simplicity and cognitive simplicity. First seeks most economical explanation while second one demands that cognitive structures must be organized in such a way that they are minimal in some sense

There are two scenarios related with simplicity. In one case we take it as the driving force for internal transformations while in second case it works as driving force behind induction from external linguistic stimuli.

Brighton’s Experiment

Taking ILM as base, Brighton did experiment to study how language evolves. He took some very weak assumptions. He assumed cultural transmission over a noiseless channel. When an agent sees a signal, he is provided the intended meaning of that signal i.e. agents have ability of “mind read”. Issues of communication are completely ignored. Each agent learns from only one agent and its performance is observed by only one agent i.e. population effects are not considered at all.

Initially we start with random mapping between meaning space and signal space. This goes as input to first generation, and then some queries are made. It performs and performance of this generation goes to next generation as input. It repeats again and again and slowly language evolves.

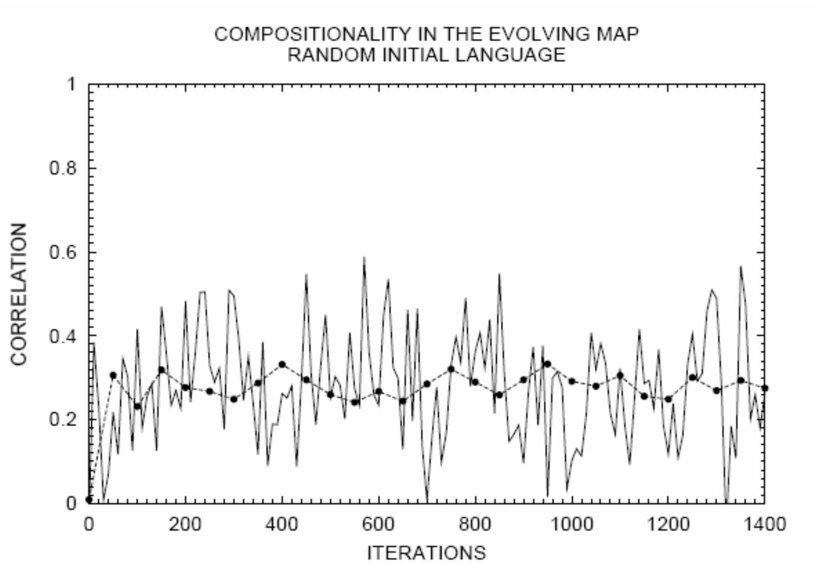


Process of language evolution: Brighton’s model

Explanatory Simplicity

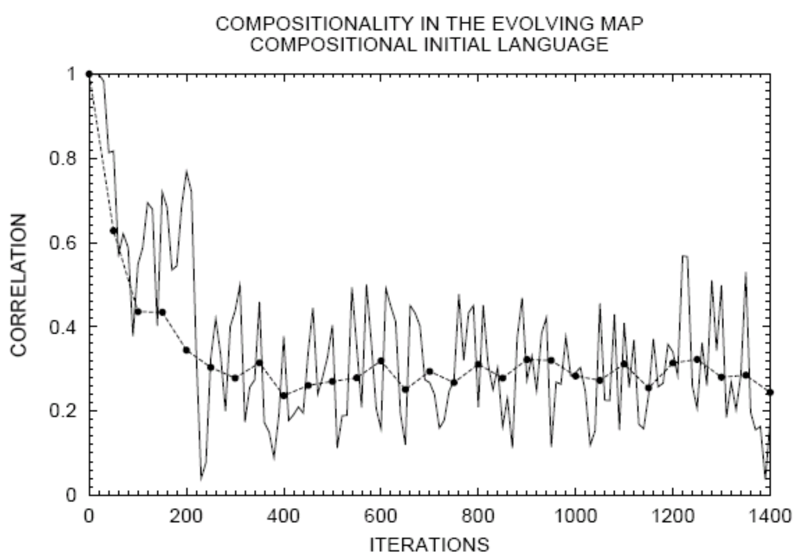
Brighton clearly used nearest neighborhood principle as a basic principle for cultural transmission and modeled it to express unobserved meanings. It allows agent to produce new signals for newly observed meanings. So this production mechanism enables agents to extend mapping taking into account its previous knowledge.

Throughout this paper correlation is assumed between experimental mappings with respect to 100% compositional language. If we start with random mapping, we end up getting correlation factor near about 0.3. We see same behavior by initially taking 100% compositional language.



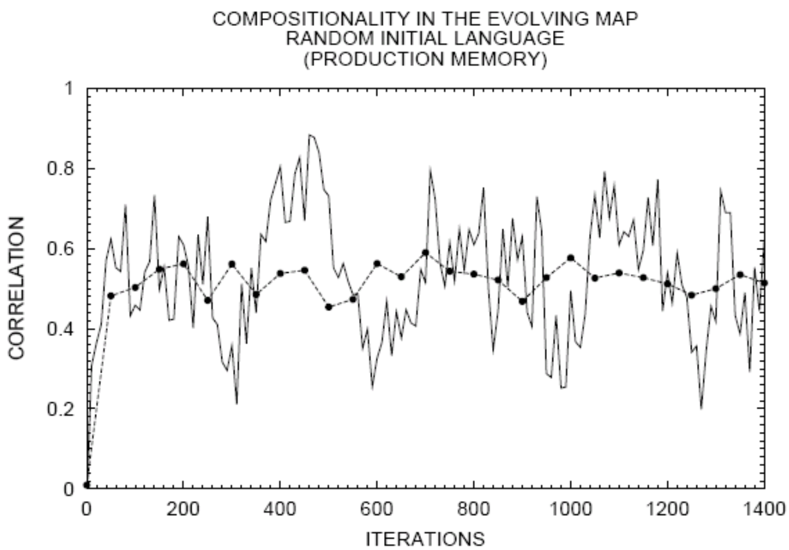
Behavior of random initial language in Brighton's experiment based on neighborhood principle

,

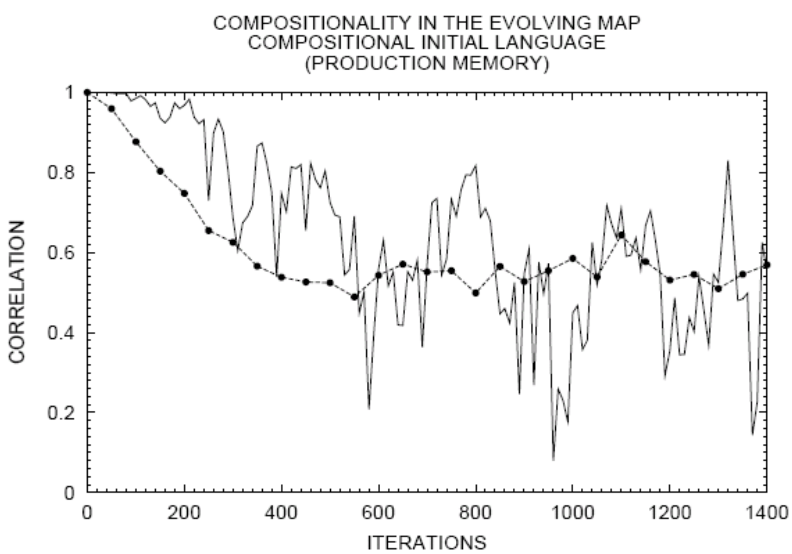


Behavior of 100% compositional initial language in Brighton's experiment based on neighborhood principle

Now, we introduce one new concept of production memory. It simply records production of an agent during its lifetime. It increases consistency between learners as well it increases coherence between adjacent meanings mapped to non-adjacent signals. Production memory very positively affects compositional property of language and correlation factor approaches to 0.6, still behavior of initial random mapping and compositional language is exactly the same.

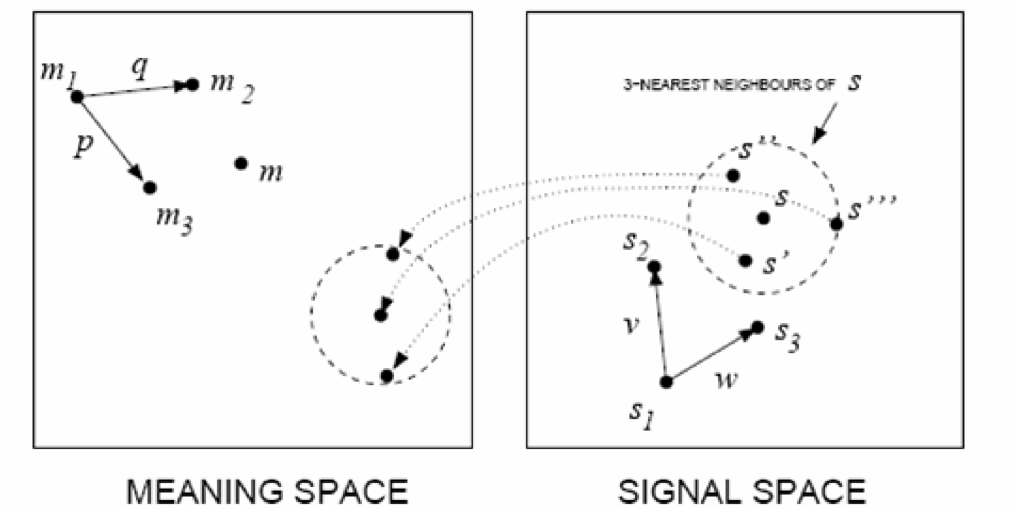


Behavior of random initial language in Brighton's experiment after introduction of production memory.

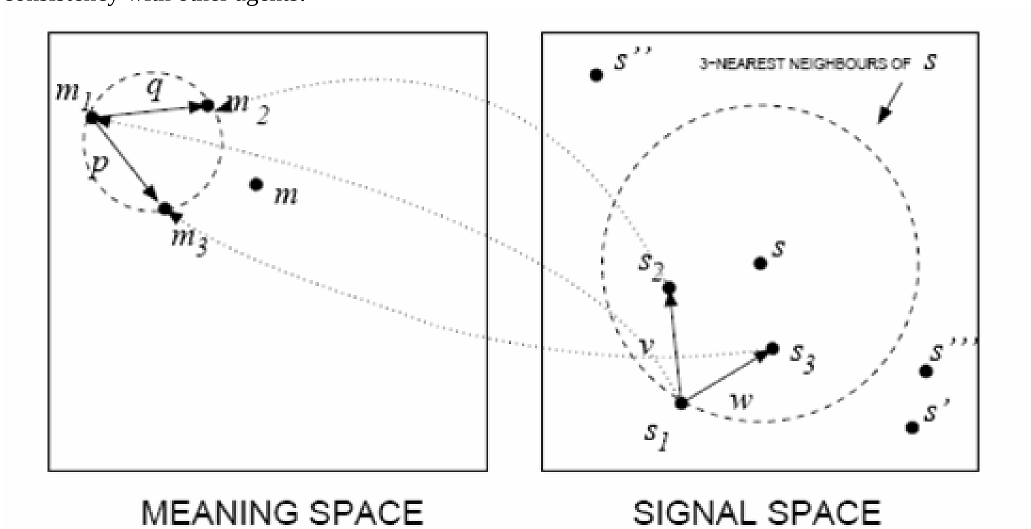


Behavior of 100% compositional initial language in Brighton's experiment after introduction of production memory.

Learner may encounter some situations in which production mechanism creates some new signals which will confuse the learner. In such situations learner rejects the signal and tries again to produce new signal. It is known as obverter procedure. This mechanism enables the learning agent to take into account of other agent's behavior. It also makes sure that neighborhood principle should not be violated and maximizes the similarity of signals used (by different agents) for same meaning. It is required because agents are not empowered with "mind read" capability all the time. They can use this capability only when they learn language from their parent.

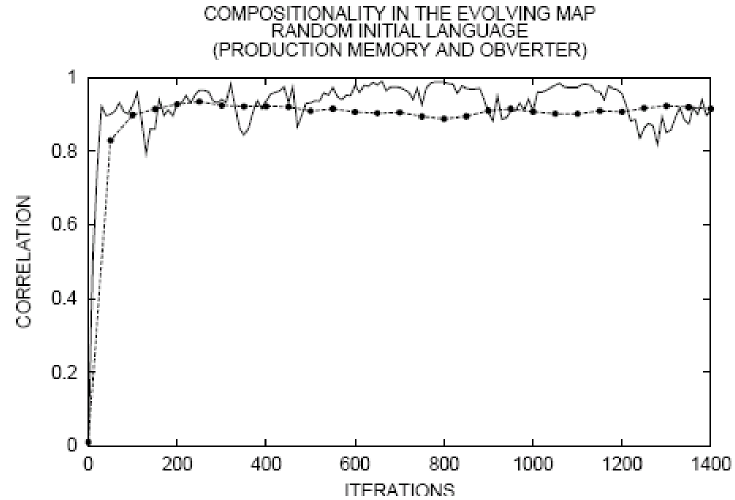


Here obverter procedure rejects production of signal ‘s’ for meaning ‘m’ as it does not show consistency with other agents.

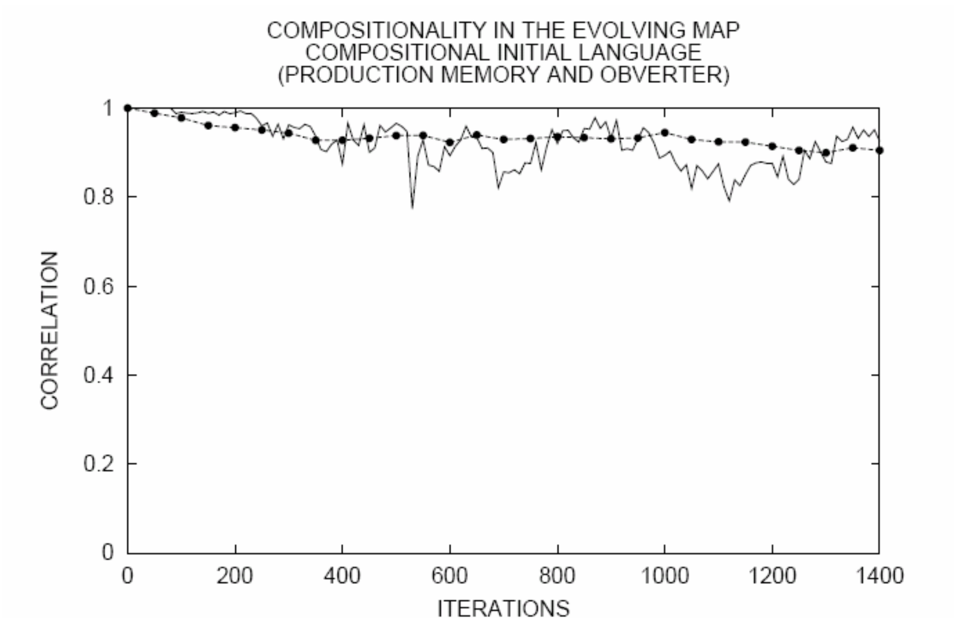


Correct choice of signal for ‘m’ taking into account other agents behavior.

Obverter principle makes a good jump towards compositionality and correlation factor approaches to 0.9, and behavior of initial random mapping and compositional language is exactly the same.



Behavior of random initial language in Brighton's experiment after introduction of production memory and obverter procedure.



Behavior of 100% compositional initial language in Brighton's experiment after introduction of production memory and obverter procedure.

.

Cognitive Simplicity

To prove cognitive simplicity we start with mathematical background. If we have a number of hypotheses to explain some phenomena, we must use the simplest. It is known as Occam's razor Principle. Rissanen proposed Minimum Description Length principle (MDL). It states that the best or simplest hypothesis for some observed data is the one that minimizes the sum of (a) the encoding length of the hypothesis, and (b), the encoding length of the data, when represented in terms of this hypothesis.

MDL principle is very useful to prove that simplicity is a guiding principal for linguistic evolution. It suggests that between two grammars always choose the simpler one.

Here we take meanings as feature vectors representing points in a meaning space.
We take two parameters to represent a meaning.
F: the number of features each meaning.
V: how many values each of these features can have.

For example if we take two features then our meaning space will be
M = {(1; 1); (1; 2); (2; 1); (2; 2)}

We will represent signals as a finite string of symbols drawn from English alphabet
S = (ba; ccad; acda; c…..)

	value 1	value 2
feature 1	a	b
feature 2	c	d
feature 3	e	f

Now three types of mapping are possible. First one is the compositional mapping in which we can draw meaning of a complex signal by understanding meaning of its constituents.

L_{Comp}={<(1,2,2), adf >, <(1,1,1), ace >, <(2,2,2), bdf >, <(2,1,1), bce >, <(1,2,1), ade >, <(1,1,2), acf >}

Second type of mapping is a completely random mapping.

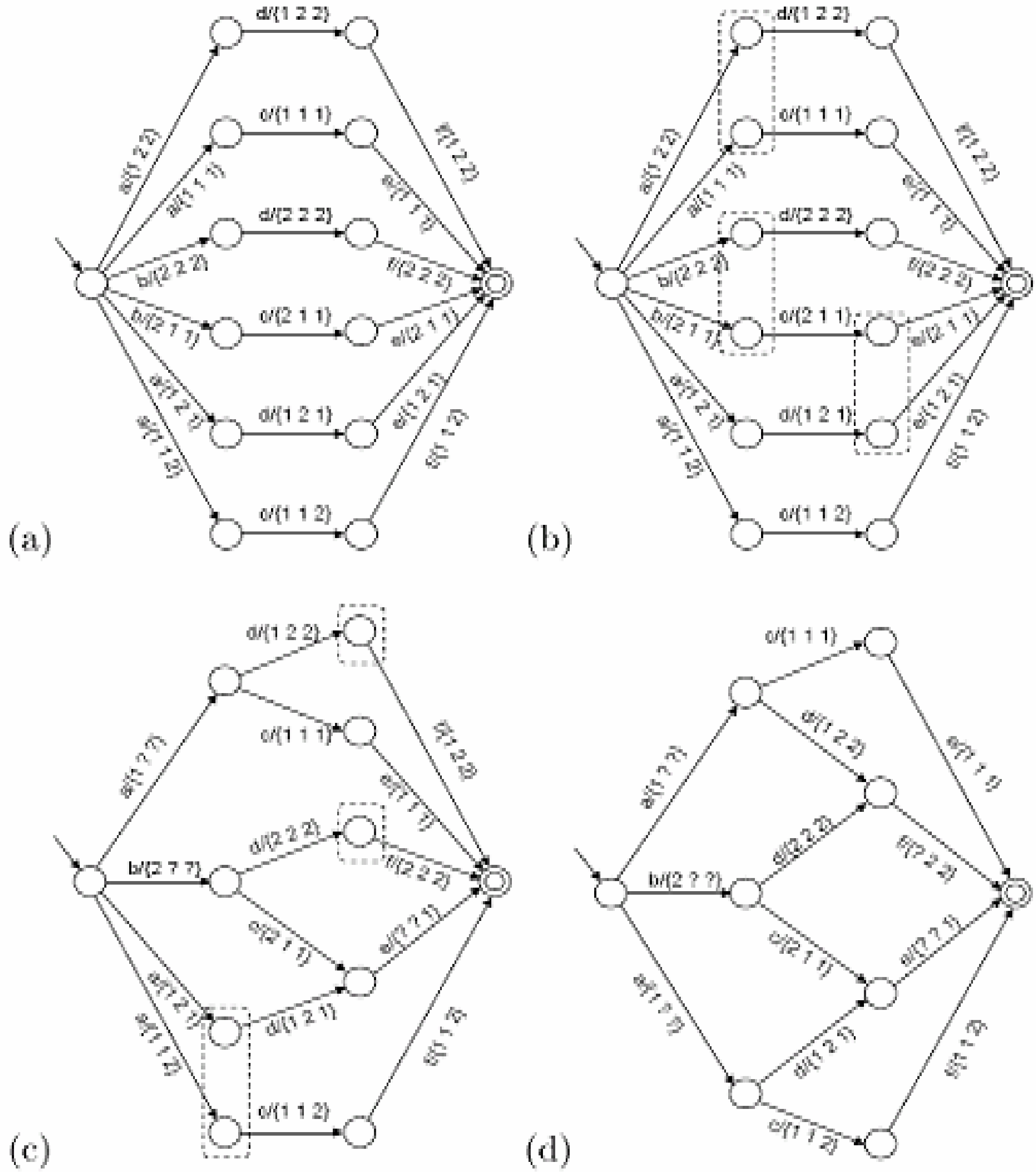
$L_{Holistic}$={<(1,2,2), sghs >, <(1,1,1), ppold >, <(2,2,2), mon >, <(2,1,1), q >, <(1,2,1), rcd >, <(1,1,2), esox >}

Third type of mapping is the mixed one.

L_{Mixed}={<(1,2,2,2), adf >, <(1,1,1,2), ace >, <(2,2,2,2), bdf >, <(2,1,1,2), bce >, <(1,2,1,2), ade >, <(1,1,2,2), acf >, <(1,2,2,1), sghs >, <(1,1,1,1), ppold >, <(2,2,2,1), mon >, <(2,1,1,1), q >, <(1,2,1,1), rcd >, <(1,1,2,1), esox >}

To encode our grammar and input data we need some encoding. For this purpose we use a finite state Unification transducer (FSUT). A finite state transducer is the one which maps one regular language to another. During this processing it attaches output symbols to each state transition within the transducer itself. We used modified form of transducer to map meanings to signals.

$$
\begin{aligned}
L = \quad & \{\langle\{1,2,2\},\mathrm{adf}\rangle, \langle\{1,1,1\},\mathrm{ace}\rangle, \langle\{2,2,2\},\mathrm{bdf}\rangle, \\
& \langle\{2,1,1\},\mathrm{bce}\rangle, \langle\{1,2,1\},\mathrm{ade}\rangle, \langle\{1,1,2\},\mathrm{acf}\rangle\}
\end{aligned}
$$

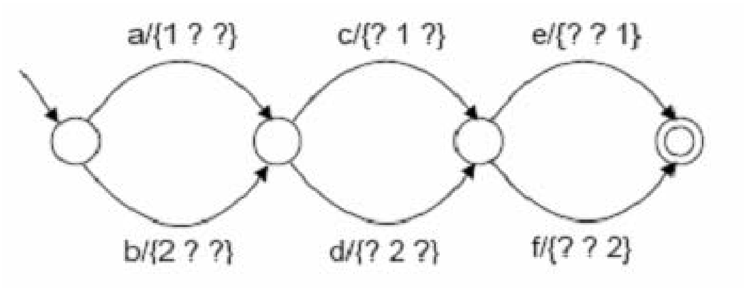


a) Figure shows representation of meaning and symbol mapping as transducers.
b) Shows Edge merge
c) Shows State merge
d) Final transducer.

Two states of transducers Q_1 and Q_2 can be merged and make a new state Q if we find that transducer is consistent during this operation. Edges going in Q_1 and Q_2 will now go into Q. Same way we can define Edge merge. Two edges can be merged, if they share same source and target state and accept same symbol. Merged edges define a new meaning label. So due to edge merging new meanings arise.

Let us take our language as following:

L_{Comp}={<(1,2,2), adf >, <(1,1,1), ace >, <(2,2,2), bdf >, <(2,1,1), bce >, <(1,2,1), ade >, <(1,1,2), acf >}

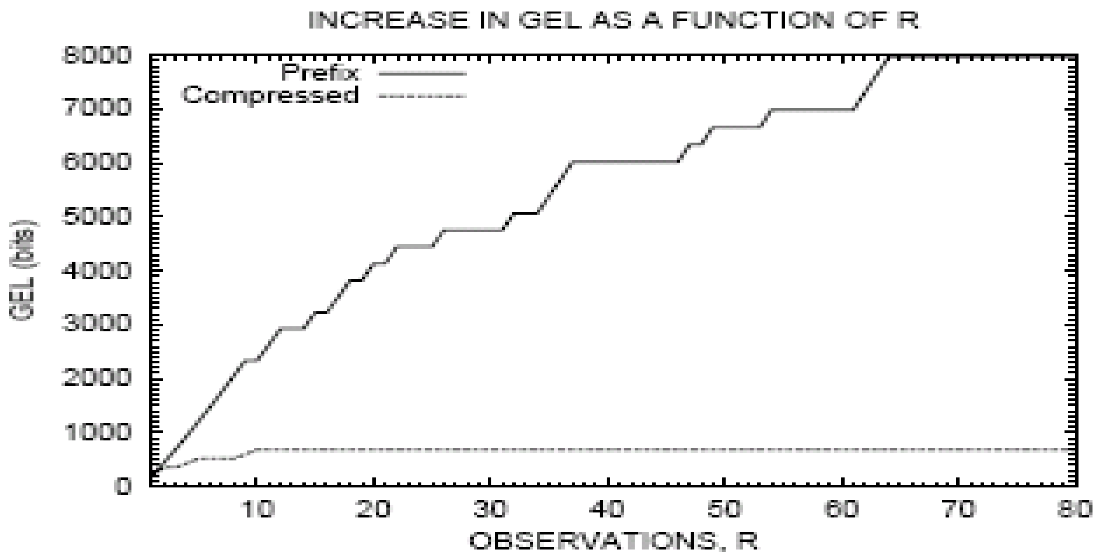


Now due to state merge and edge merge we will get

L^{+}_{Comp}={<(1,2,2), adf >, <(1,1,1), ace >, <(2,2,2), bdf >, <(2,1,1), bce >, <(1,2,1), ade >, <(1,1,2), acf >, <(212), bcf>, <(2,2,1), bde>}

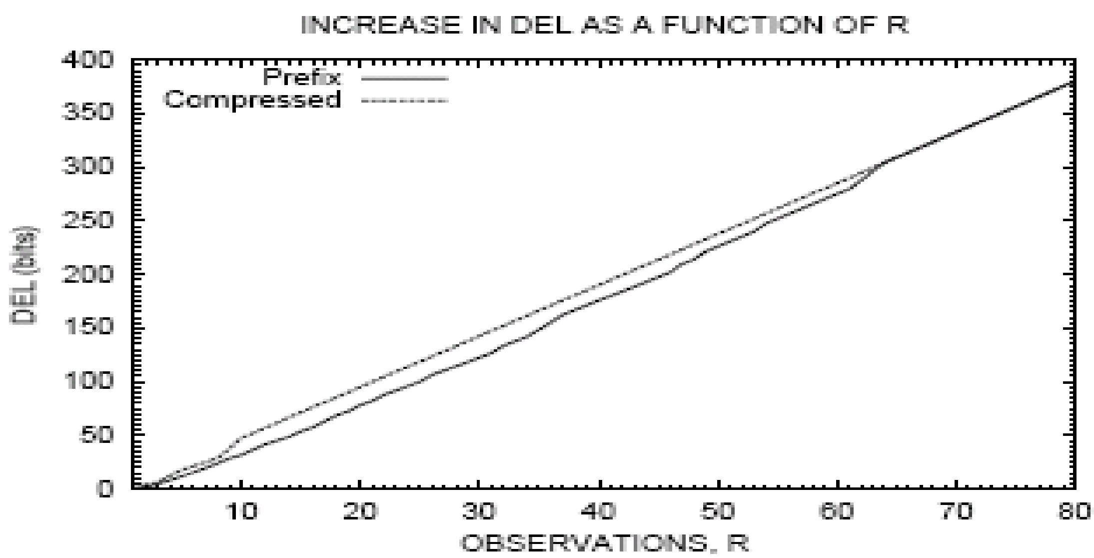
But these meanings also preserve compositionality property. So here we don't see any contradiction.

Brighton proved that encoding based on transducer will be minimal. We take his statement as conjecture and would like to avoid a long mathematical proof.
As number of observations increases, bits needed to represent Length of encoded Grammar (GEL) increases if we are using prefixed grammar. Yet if we use FUST encoding to represent the grammar we see that GEL is almost constant after some finite iterations.



Comparison between GEL of prefixed grammar and compressed grammar.

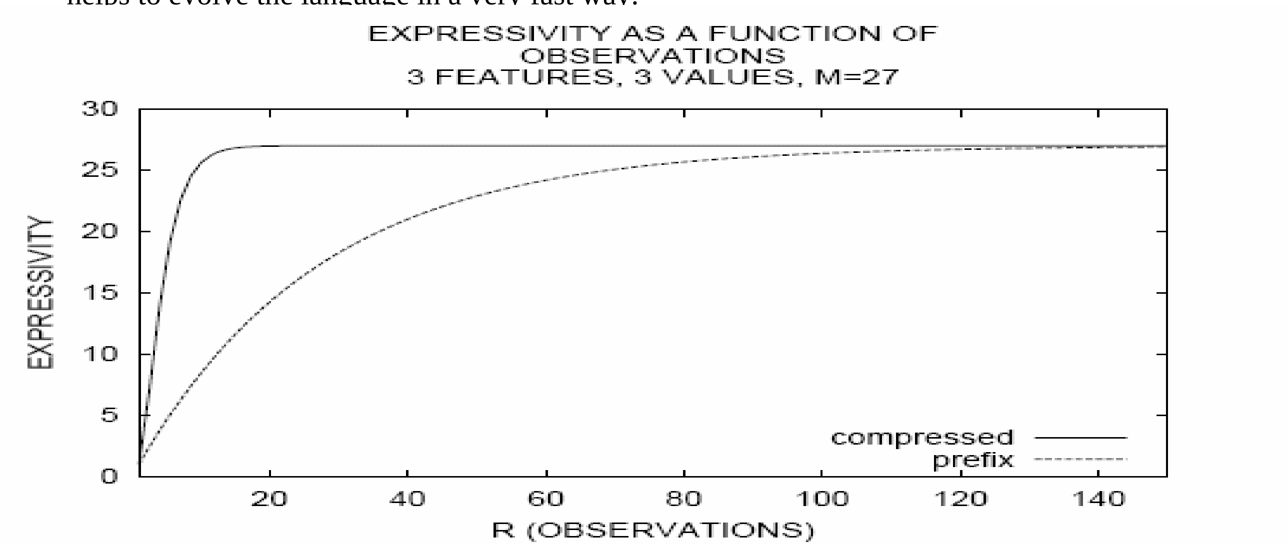
In case of Data Input Length (DEL) compression hardly matters. We see minute difference between prefixed DEL and compressed DEL which disappears after reasonably finite number of iterations. So for our study, we can take GEL as benchmark.



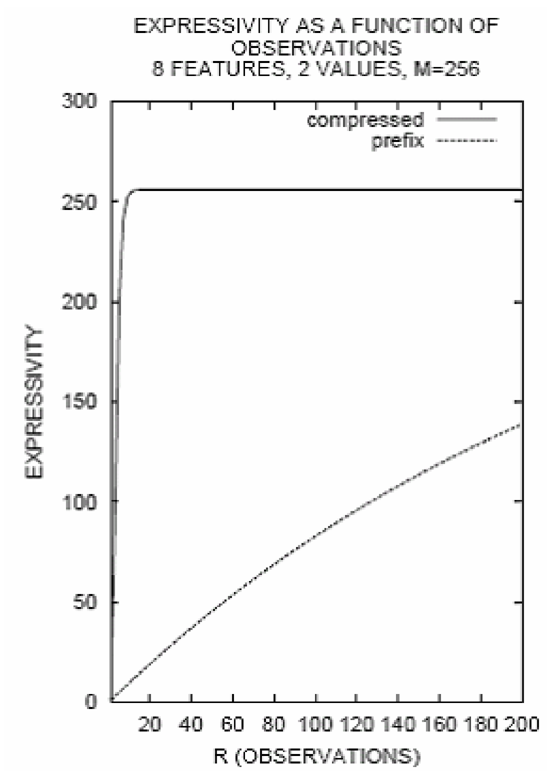
We here assumed all points of meaning space as equi-probable. To make our assumption more realistic, we assume that all meanings are not observed in a given generation.

When we start from a random mapping, finally we end up getting few meanings mapped to signals in a particular generation. This clearly shows that only these meanings can be used by that generation. Hence it is considered as the most important factor in study of language evolution. Formally we call it Expressivity of the language. Hence Expressivity is a measure of the number of meanings (out of total possible meanings) for which a signal can be constructed.

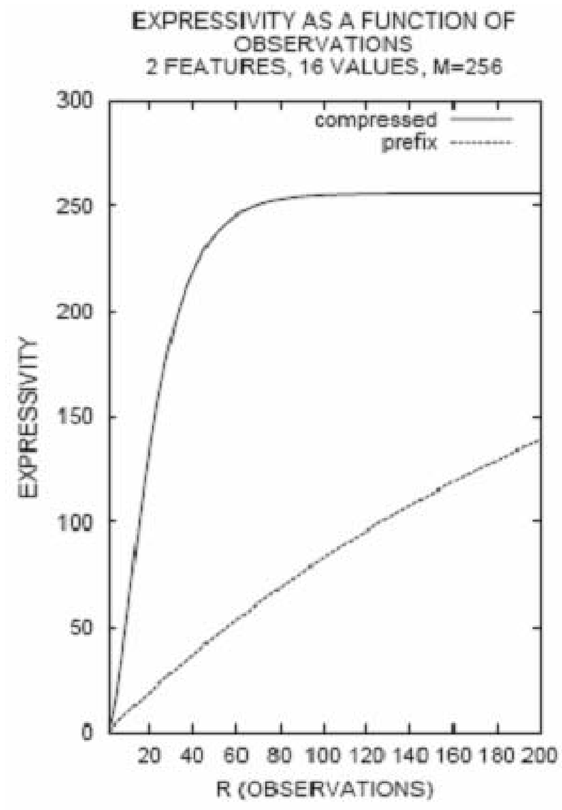
We take number of features of each meaning and values (that each of these features can take) as input. We study behavior of language during transmission in terms of expressivity. More expressivity means better evolution of language. We encode grammar using FSUT encoding and compare it with other prefixed encoding it terms of expressivity. In Every experiment, we observe that compressed grammar turns out to more expressive in finite number of iterations and it converges very fast. It simply shows that simplicity is a driving force behind language evolution because it helps to evolve the language in a very fast way.



Expressivity of compressed grammar language goes up very fast but prefixed one also approaches it after a number of iterations



Here we see that compressed reaches to 100% expressivity within a few number of iterations while prefixed one is far behind.



Here we see that compressed language reaches to 100% expressivity within a few number of iterations while prefixed one is far behind. Still it takes more iterations than last example.

Key Points

Explanation of Hallmarks of languages is taken as main basis for study. Here we clearly claim why human languages are different from animal languages. Transmission between generations is explained and it's really a good model to represent language evolution due to transfer of knowledge of language from one generation to another. We have simplicity as a motivation for linguistic evolution and it emphasizes good old principle—Simpler is better and faster.

Main Drawbacks

Which came first I-language or E-language? It is like good old egg and chicken problem because we considered that I-language transform into E-language and then E-language transforms into I-language. But from where we started – I-language or E-language?? Interaction within a generation is not taken care here because one learner learns from only one agent. Also interaction of agent with other previous generation members is also not taken into account. We proved that Simplicity is a driving force for language evolution but we have no Idea that it is the only reason for language evolution or there are many more?? We also failed to take into account intentional and ambiguous sentences.

Conclusion

Whole analysis is based on sound mathematical foundation and logical reasoning. Results of simulation are good and show the importance of simplicity as driving force in language evolution still we must always remember that the proposed model is not based on real life experiments. We can't do our experiment on animals and humans have a very long life-span. Hence we must do some experiments with bees because their dance is compositional and their life span is short. So we can easily do our experiments and see how cultural transmission happened. Without any such experiment this model can't we taken as real evolutionary model and would remain a mathematical model only.

Reference:

1. BATALI, John.1998. *Computational simulations of the emergence of grammar in approaches to the evolution of language: social and cognitive bases*, Cambridge University Press

2. BRIGHTON, Henry. 2002. *Compositional syntax from cultural transmission*. Artificial Life, the MIT Press.

3. BRIGHTON, Henry. 2003. *Simplicity as a Driving Force in Linguistic Evolution.* The University of Edinburgh.

4. KIRBY, Simon. 1999. *Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners*. Cambridge University Press.