Simplicity as a Driving Force for Language Evolution

Ram Ji Tripathi
B. Tech (CSE)
Y1284

Project Guide
Prof. Harish Karnick
Prof. Achla M. Raina

Language Evolution: Overview

- How did language come to have its characteristic structure?
  - the hallmarks of language are simply properties of our biological machinery
  - language is learned
- the conditions for linguistic evolution towards compositional structure correspond to
  - specific levels of semantic complexity, and
  - induction based on sparse language exposure.

Language is Innate??

- It is hard to avoid the conclusion that a part of the human biological endowment is a specialized “language organ," the faculty of language (FL). Its initial state is an expression of the genes, comparable to the initial state of the human visual system, and it appears to be a common human possession to close approximation. --Chomsky
- The poverty of the stimulus, is a conjecture stipulating that the knowledge of language children attain is surprising precisely because it cannot be derived solely from information made available by the environment.-- (Chomsky & Wexler).
- If knowledge of language is innate, then why does language exhibit so much variation?
- abstract properties of language are encoded in the genes

Language is Innate??

- languages themselves adapt as a result of being repeatedly transmitted from one generation to another through the processes of production and induction
- rather than the hallmarks of language residing in biological structure that is an expression of the genes, they are taken to be artifacts reflecting the accumulated residue of language transmission
- language universals are artifacts of a little-understood interaction between a cognitive system and an environment of adaptation

Language Evolution

- The process by which languages themselves evolve is termed linguistic evolution
- on the basis of universal features of language, we can propose a universal grammar (UG)
  - UG can be employed as the backbone for a predictive theory of languages not just observed, but also for those languages that are possible
  - UG is taken as a plausible object to be explained by the cognitive sciences.

Language Evolution

- I agree that learning which makes more effective use of the input certainly helps the child, and it certainly takes some of the load of the Universal Grammar. But I do not think it takes all the load . It may allow Universal Grammar to be less rich, but it does not allow UG to be dispensed with altogether. ---Jackendo

The principle of detachment

- Explanation of the cognitive processes relevant to language, coupled with an understanding of how these processes mediate between input (primary linguistic data) and output (knowledge of language), would be sufficient for a thorough explanation of the universal properties of language. In other words, an explanation for the universal features of language amounts to an explanation of the possible states of the biological machinery elicited by the linguistic input. –brighton
- The principle of detachment suggests that the biological machinery underlying the processing and understanding of language is the principal object of study.

Evolution

- The communication systems used by animals do not even approach the sophistication of human language, so the evolution of language must concern the evolution of the human line over the past 5 million years
- language evolved in humans due to the functional advantages gained by linguistically competent humans

Evolution

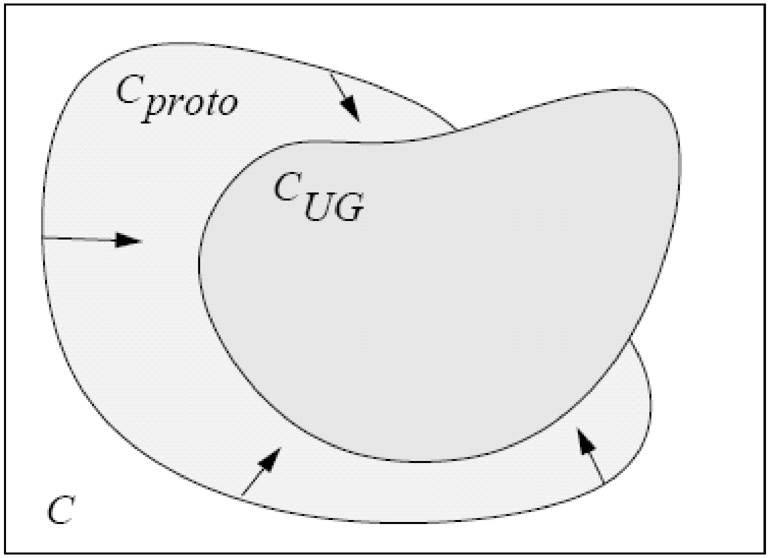


Figure 2.1: Functional nativism. From the set of all communication systems C, the communication systems of proto-humans, C_{proto}, evolved under some functional pressure towards C_{UG}.

Iterated Learning

- biological evolution relies on a genetic substrate: information relevant to language is encoded in the genes.
- cultural transmission and adaptation allows information relevant to language to be encoded within languages themselves
- An iterated learning model (ILM) is a framework for testing theories of linguistic evolution. Within an ILM agents act as a conduit for an evolving language

Iterated Learning

- the language itself changes or evolves rather than the agents themselves.
- An ILM is a generational model: after members of one generation learn a language, their production becomes the input to learning in the next generation.
- Within this model of linguistic transmission, providing that the transfer of knowledge of language from one generation to the next is not entirely accurate or reliable, will result in diachronic change.
  - certain linguistic structure will survive transmission, while other forms may disappear
- Language evolution

Iterated Learning

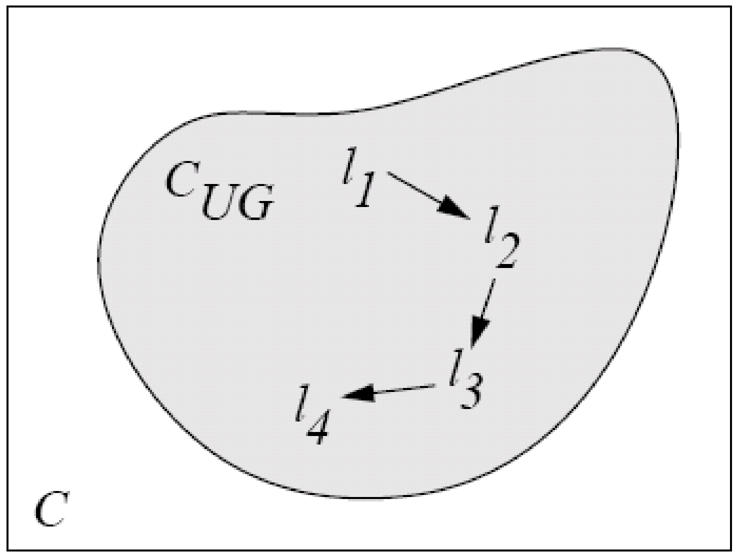
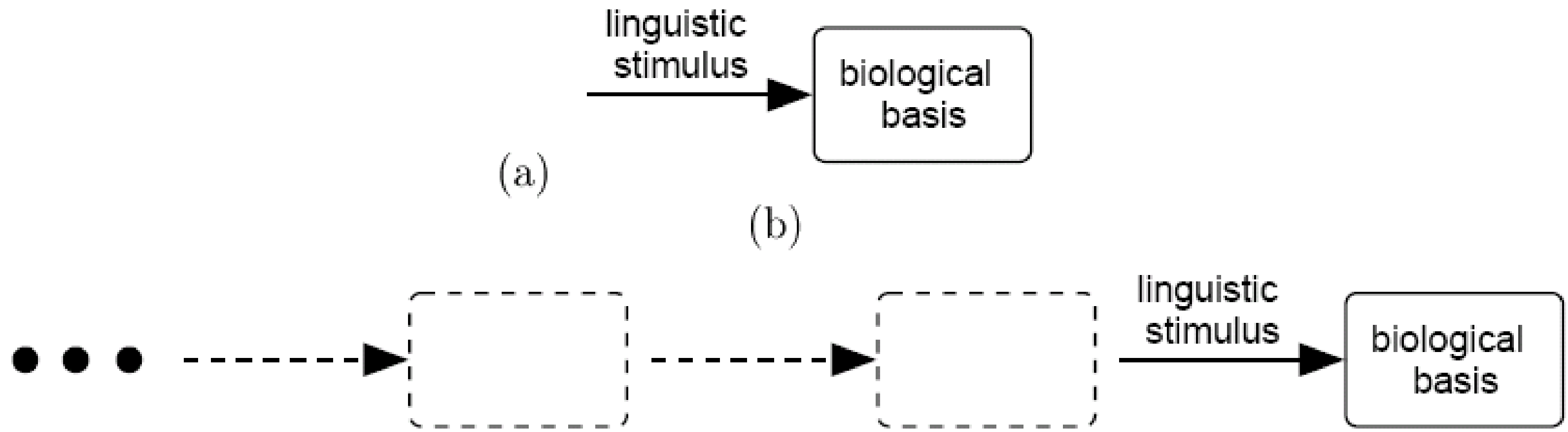


Figure 2.2: Language change. An example trajectory of language change through languages l_1, l_2, l_3, and l_4.

Iterated Learning

- Cultural adaptation: The occurrence of changes in the language due to the effects of cultural transmission.
- Cultural Selection for Learnability : In order for linguistic forms to persist from one generation to the next, they must repeatedly survive the processes of expression and induction. That is, the output of one generation must be successfully learned by the next if these linguistic forms are to survive.
- Those forms that repeatedly survive cultural transmission are adaptive in the context of cultural transmission: they will be selected for due to the combined pressures of cultural transmission and learning.

Iterated Learning



Iterated Learning

- I-language is the internal configuration of cognitive structures relevant to knowledge of language.
- E-Language is the externalized linguistic performance derived from internal linguistic competence, that is, I-language.
- Iterated learning therefore hinges on the observation that linguistic transmission is meditated by repeated translations between I-Language and E-Language.

Hypothesis

- Innateness: Humans must have a biologically determined set of predispositions that impact on our ability to learn and produce language.
- Situatedness hypothesis: A thorough understanding of the cognitive basis for language would not amount to a total explanation of universal language structure. However, a thorough understanding of the cognitive basis for language in conjunction with an understanding of the trajectory of language adaptation through cultural transmission would amount to a total explanation of language structure.
- Function independence hypothesis : Some aspects of language structure can be explained independently of language function.

Iterated Learning

- The process of cumulative cultural evolution requires not only
  - creative invention but also, and just as importantly,
  - faithful social transmission

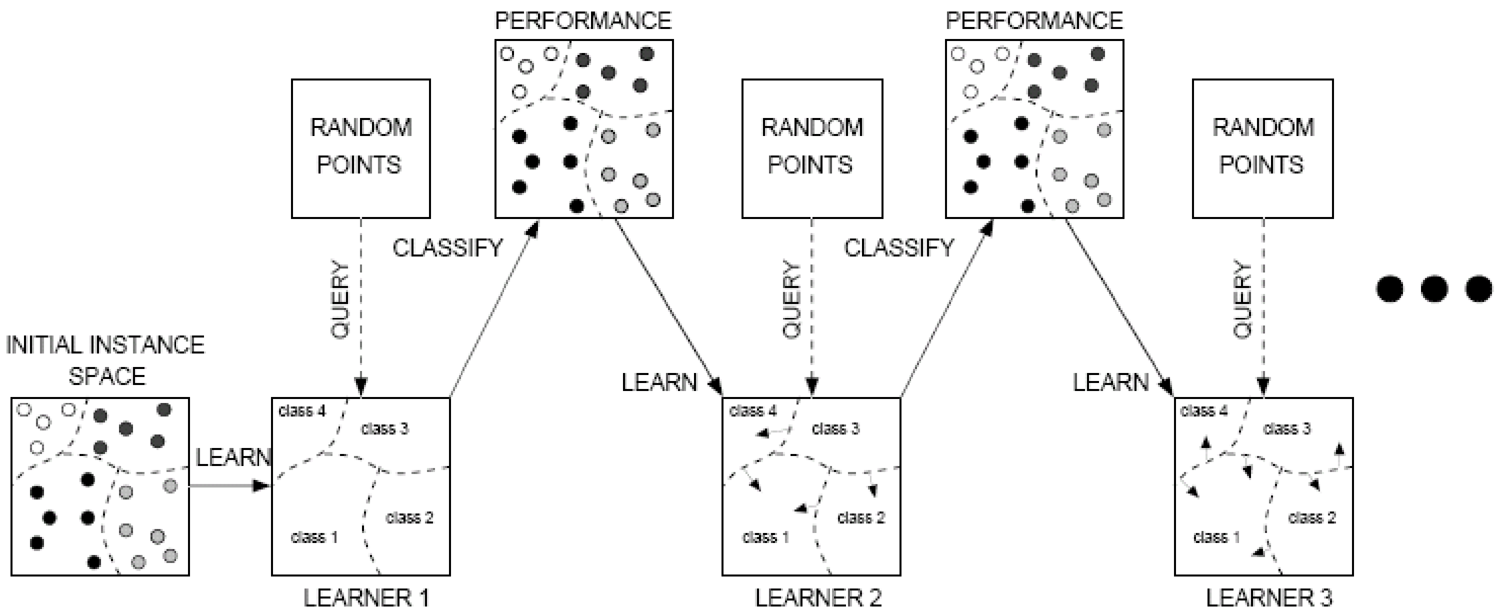
Batali’s Model

- focus on the emergence of compositionality
- Batali's model uses recurrent neural networks to map signals to meanings. In a fixed population without any turnover,
- Batali shows how repeated inter-agent production and learning of meaning form pairs results in structured signals emerging..
- conclusion is that such rudimentary syntactic structure, can be explained through repeated linguistic expression and induction.

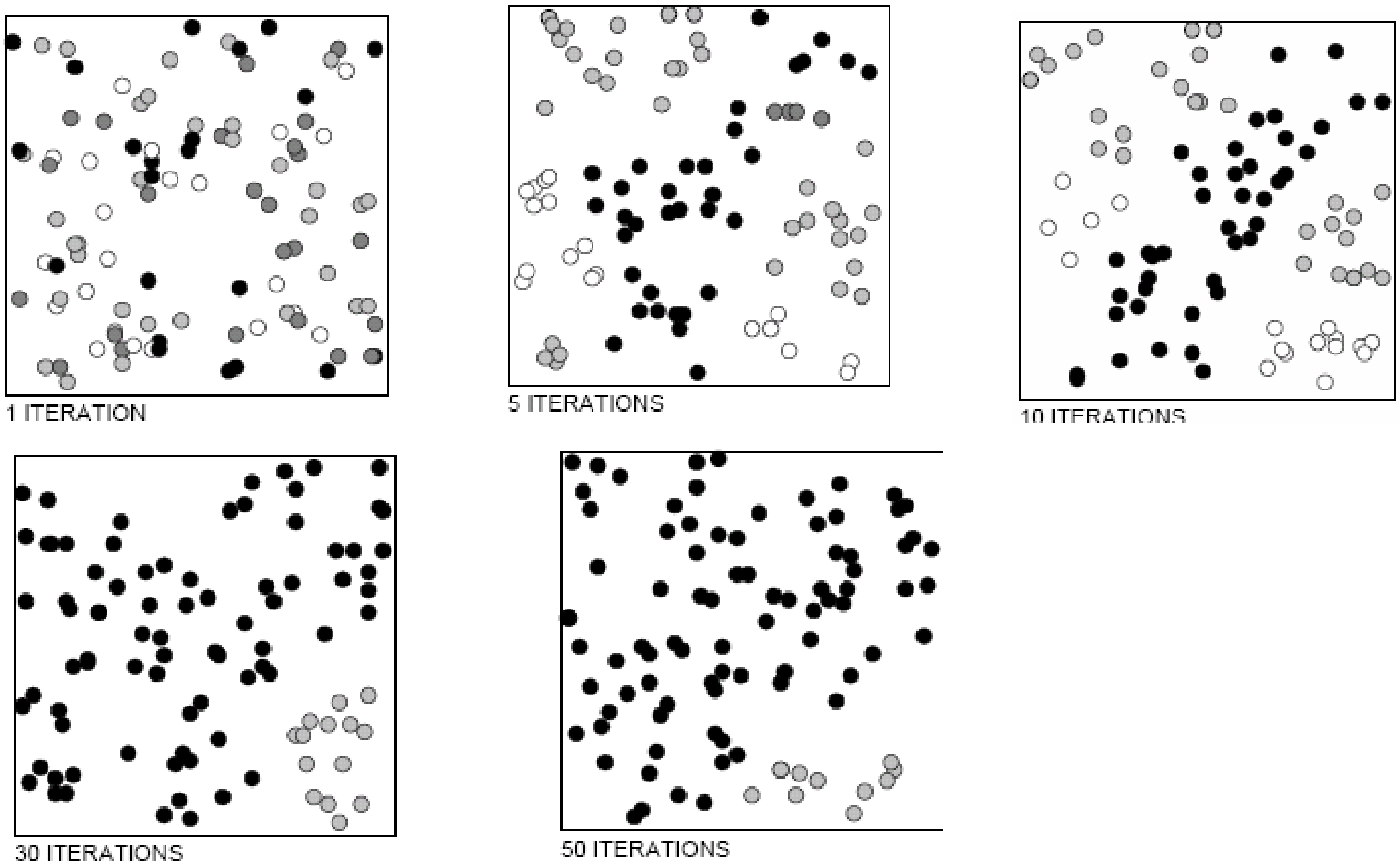
Kirby's Model

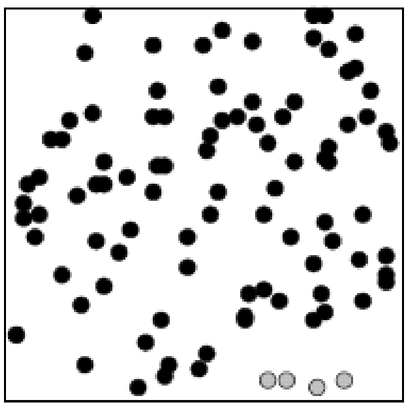
- Kirby models a collection of agents subject to a turnover.
- After a number of rounds of communication, a random agent is removed and replaced by a new blank agent..
- This transmission is explicitly presented in terms of the translation between I-language and E-language.
- The outcome of the model is that, from an initially holistic relation between meanings and signals, a compositional relation emerges as the result of the evolution of the signals.
- Both these models therefore make a claim about the emergence of compositionality from initially holistic systems. The driving force for this linguistic evolution is iterated learning: learners learn from the output of other learners

Iterated Learning

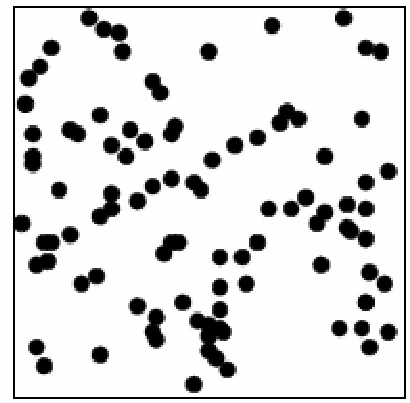


Iterated Learning





59 ITERATIONS



60 ITERATIONS

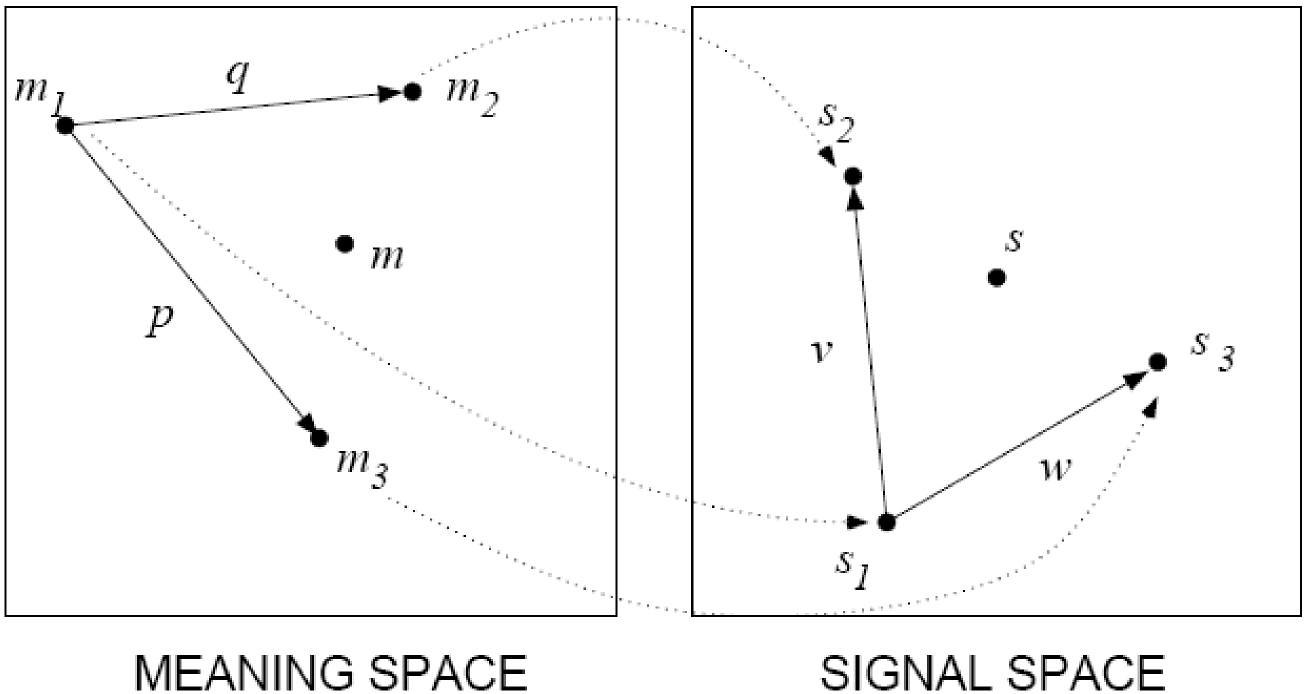
Language

- Language as a structured mapping. We can imagine language as an infinitely large mapping from meanings to signals.
- The competence of each learner represents a mapping between an infinitely large meaning space and an infinitely large signal space.
- Human languages are compositional: the mapping from meanings to signals is structured in such a way that the meaning of a signal is a function of the meaning of its constituent parts
  - compositional language will have neighborhood preserving properties. This means that similar meanings will map to similar signals.
  - because these meaning components are common to both meanings, parts of their corresponding signals must also share components. The signals will occupy neighboring parts of the signal space, just as the meanings do.

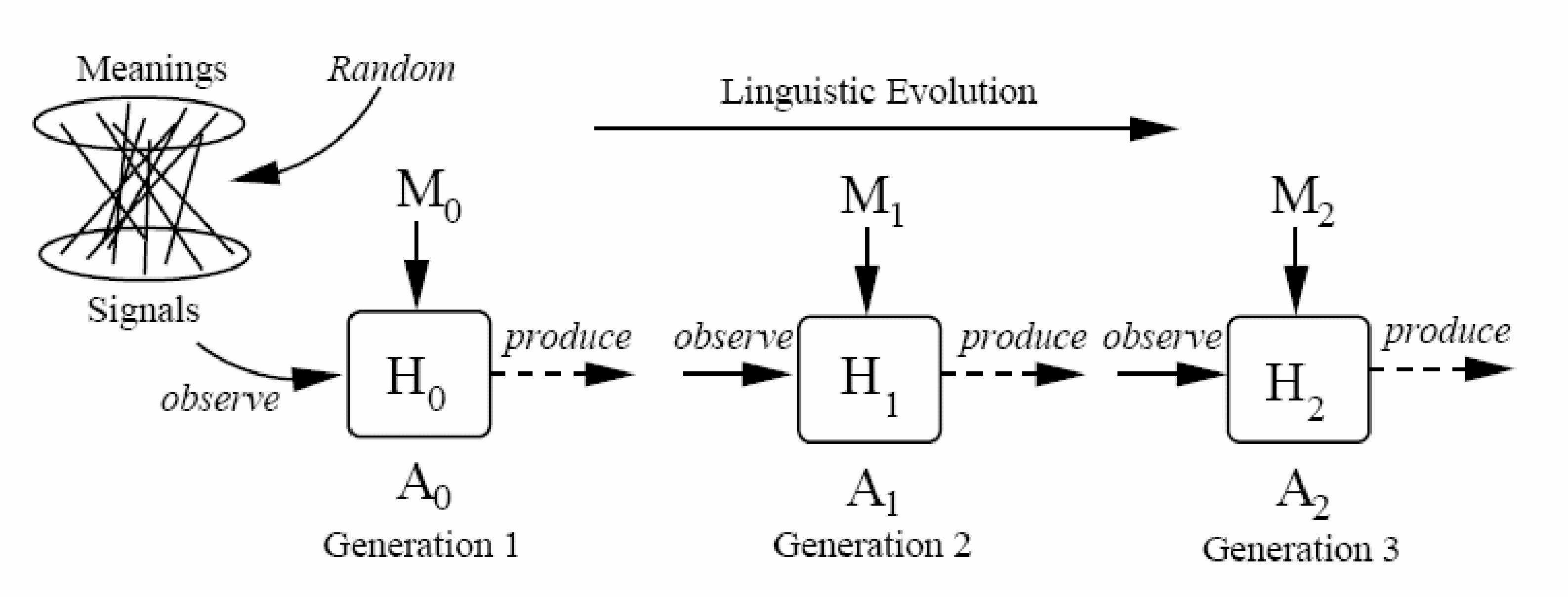
Assumptions for Brighton Experiment

- Agents in the model have the ability to "mind read". When an agent observes a signal, the intended meaning of that signal is also provided i.e.meaning transmission occurs over a noiseless channel.
- Issues of communication are not considered. For example, communicative accuracy, the agents intentions, or any model of success or failure in language use is not considered.
- Population effects are not considered. Each agent learns from the output of only one other agent. Similarly, utterances produced by an agent are only observed by a single learner.

Brighton Experiment



Brighton Experiment

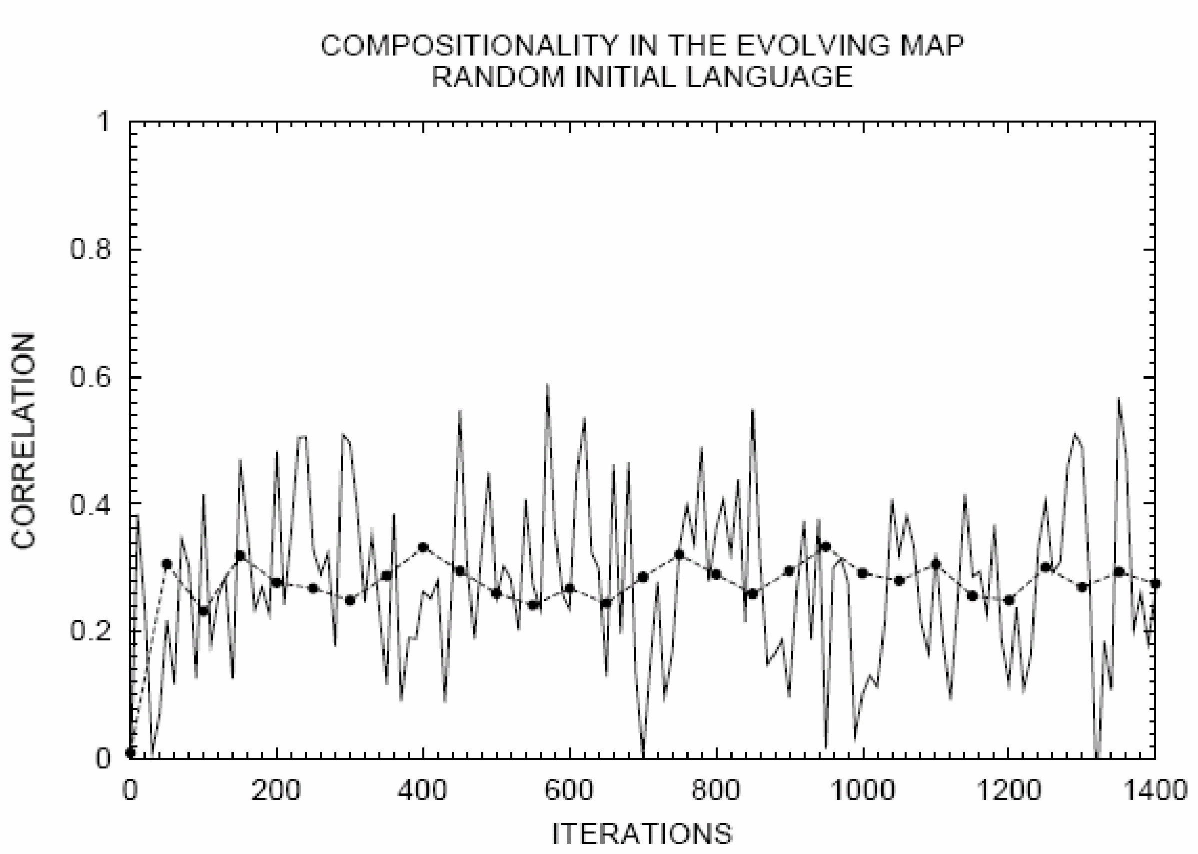


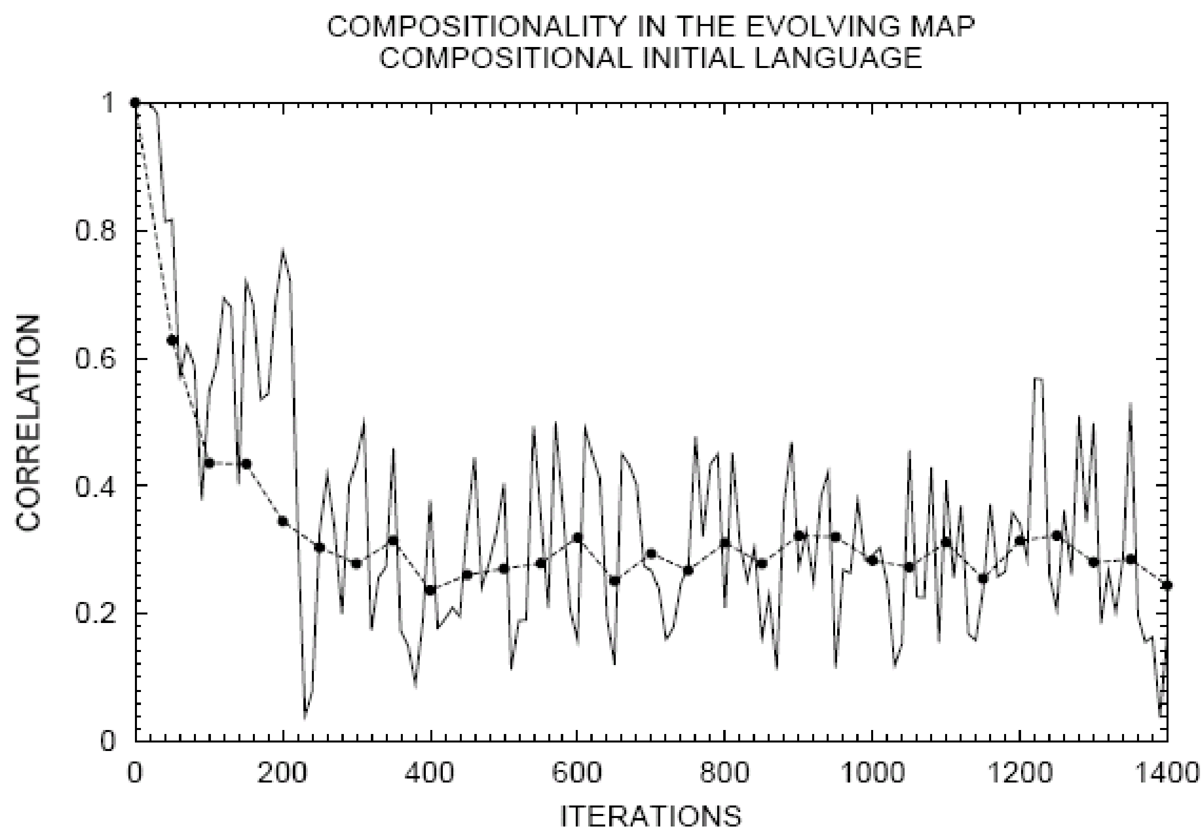
Brighton Experiment

- Nearest-Neighbour Decision Rule. Generalisation, which enables unobserved meanings to be expressed, is achieved by recalling similar previously observed instances of production.
- Production Mechanism. The production mechanism takes the relevant information found by the decision rule and uses it to synthesise a new signal. The production scheme assumes that nearby production decisions should inform a novel production decision.
- Production Memory: The production memory records the production behaviour of an individual over its lifetime. During an agent's lifetime, production decisions will therefore be partially informed by previous production behaviour.
- Obverter Procedure: The obverter procedure introduces a reject option for the process of production. If the production process results in a signal which would be incorrectly interpreted by the agent itself, then, rather than proceeding with this production behaviour, the learner instead declines to produce.

Results

- However long the system is allowed to cycle through the twin processes of production and induction, the learning and production biases alone cannot un-tangle the messy relationship between meanings and signals
- An arbitrary meaning m has three nearest neighbours. In the perfectly compositional initial system, these neighbours map to exactly the same points in the signal space. Providing the relation between m and its neighbours is faultlessly parallelled in the signal space, as it should be, production for m should lead to a point in the signal space with exactly the same coordinates as m itself. This, however, does not happen.
  - the arithmetic operations carried out by the production mechanism can introduce very slight arithmetic errors.

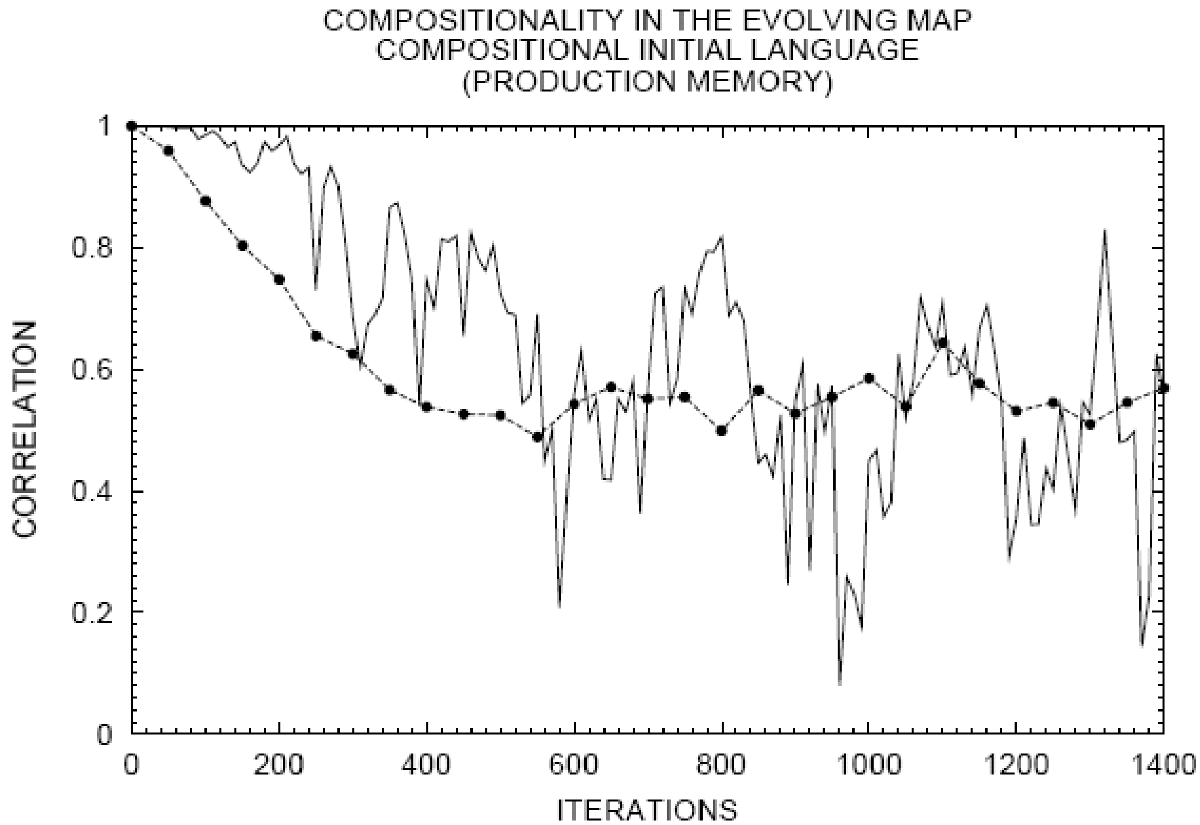

COMPOSITIONALITY IN THE EVOLVING MAP
RANDOM INITIAL LANGUAGE
CORRELATION
ITERATIONS
0
0.2
0.4
0.6
0.8
1
0
200
400
600
800
1000
1200
1400


COMPOSITIONALITY IN THE EVOLVING MAP
COMPOSITIONAL INITIAL LANGUAGE
CORRELATION
ITERATIONS
1
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000
1200
1400

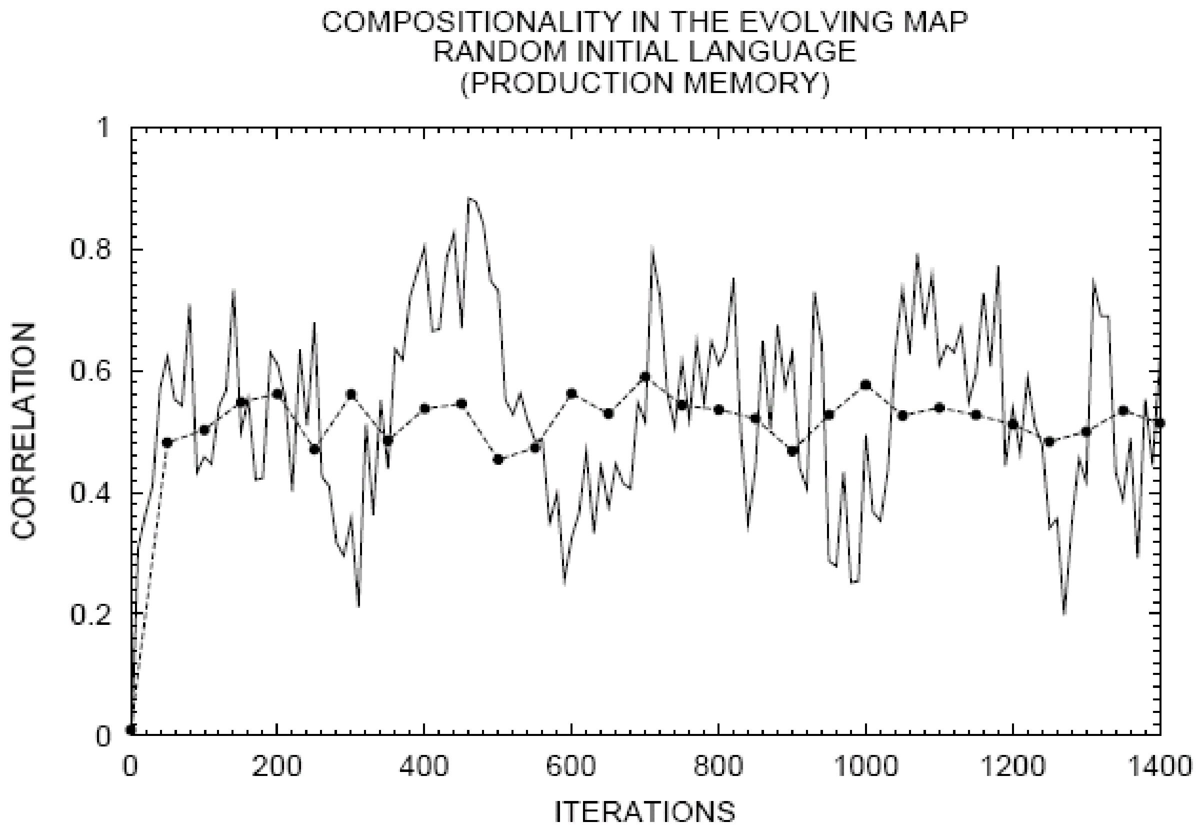
Brighton Experiment

- The production memory requires that learners update their own competence in light of production decisions they make over their lifetime. The production memory
  - smoothing the differences between adjacent regions of the meaning space that map to non-adjacent regions of the signal space,
  - increasing consistency between learners.

Brighton Experiment



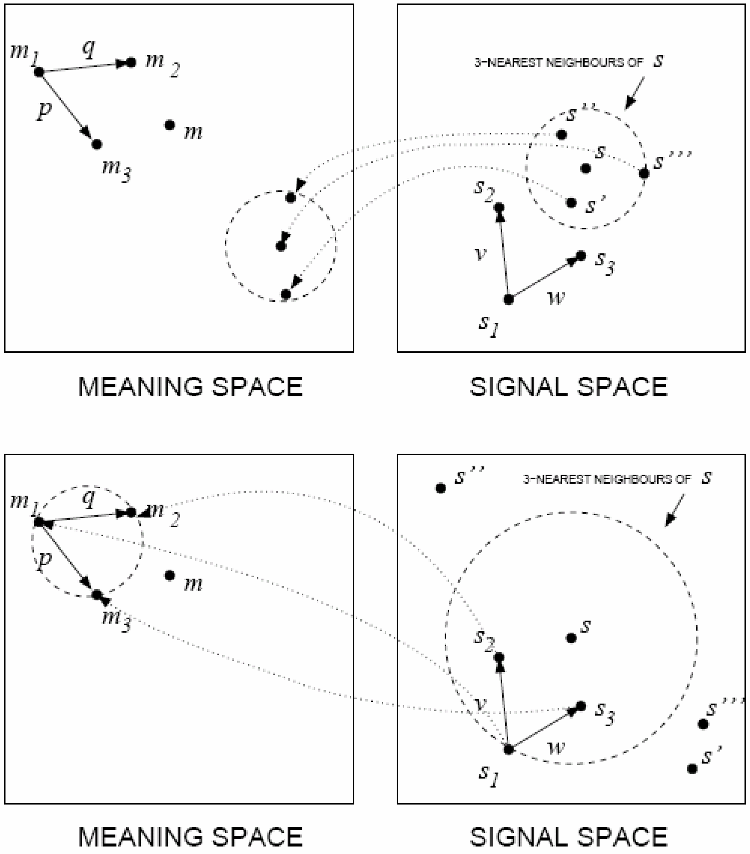
Brighton Experiment



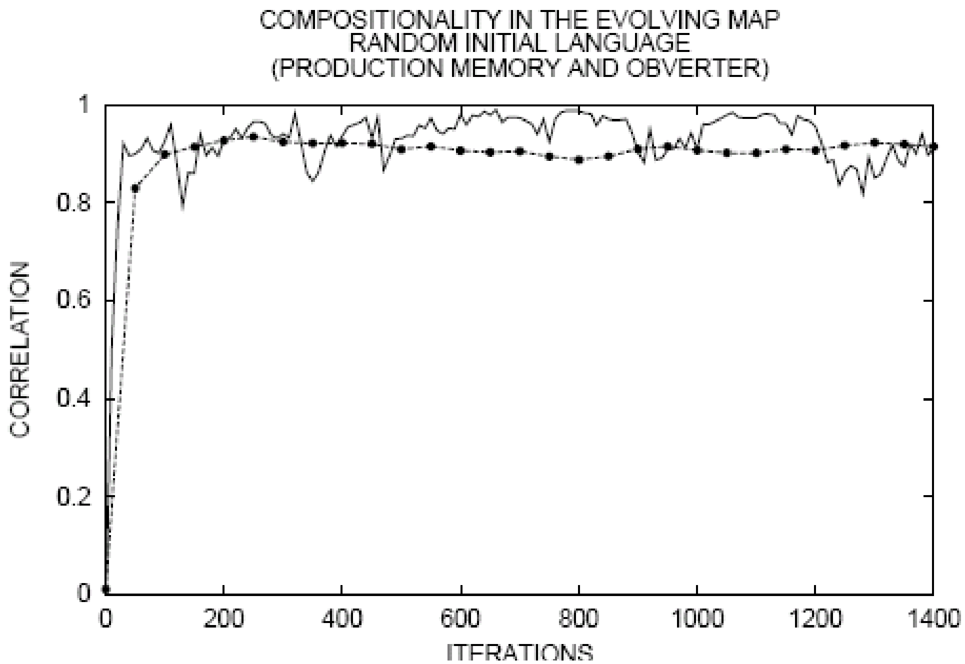
Brighton Experiment

- The obverter procedure requires a learner to model its mapping from meanings to signals on the basis of the reception behaviour of the surrounding population.
- This reception behaviour represents the manner in which other agents map signals to meanings. For example, to find an appropriate signal for a meaning m, the obverter procedure would use a knowledge of which signal leads others to infer that the meaning m has been transmitted.
- An obverter procedure informs the process of production by maximising the likelihood of the proposed signal correctly being interpreted as carrying the intended meaning

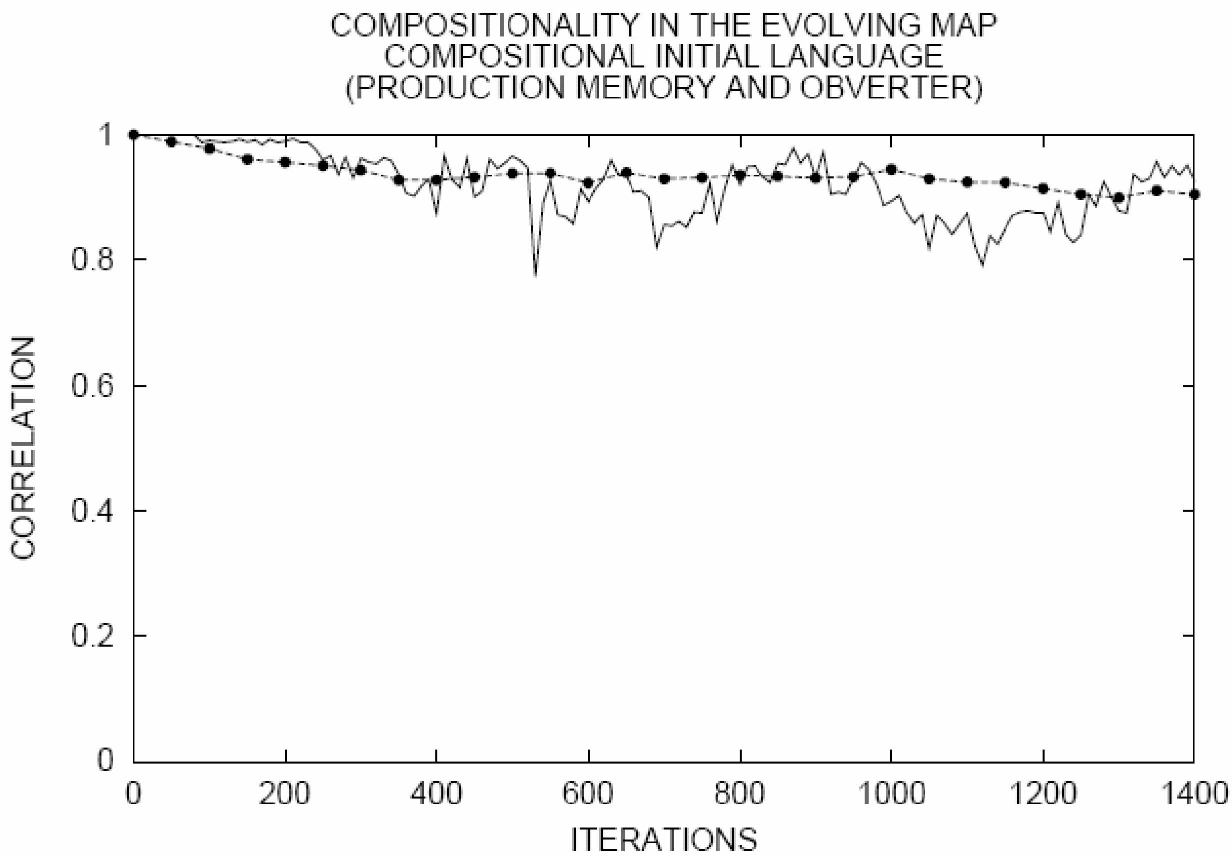
Brighton Experiment



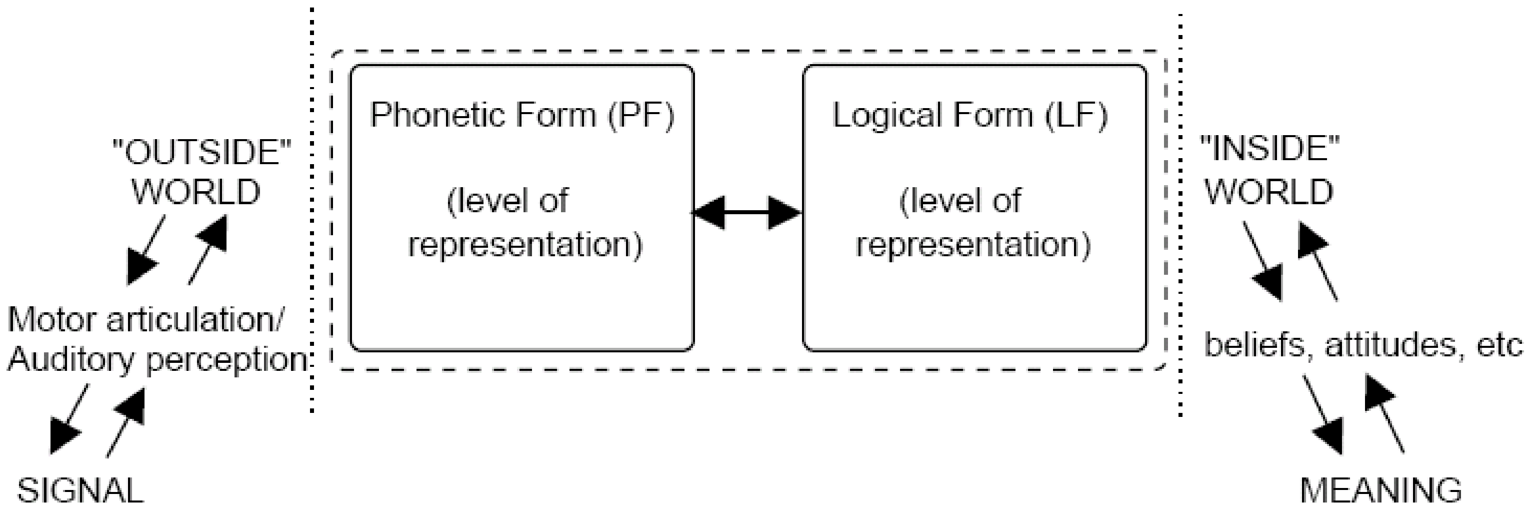
Brighton Experiment



Brighton Experiment



Brighton Experiment



Language as a Mapping

- Meanings are defined as feature vectors representing points in a meaning space.
- two parameters, F and V .
  - F :the number of features each meaning
  - V :how many values each of these features can
- M = {(1; 1); (1; 2); (2; 1); (2; 2)}
- Signals are represented as a finite string of symbols drawn from some alphabet
- S = (ba; ccad; acda; c…..)

Language as a Mapping

$$L_{compositional} = \{\langle\{1,2,2\}, \mathsf{adf}\rangle, \langle\{1,1,1\}, \mathsf{ace}\rangle, \langle\{2,2,2\}, \mathsf{bdf}\rangle, \langle\{2,1,1\}, \mathsf{bce}\rangle, \langle\{1,2,1\}, \mathsf{ade}\rangle, \langle\{1,1,2\}, \mathsf{acf}\rangle\}$$

•Those languages with no compositional structure whatsoever I will term holistic languages

$$L_{holistic} = \{\langle\{1,2,2\}, \mathsf{sghs}\rangle, \langle\{1,1,1\}, \mathsf{ppold}\rangle, \langle\{2,2,2\}, \mathsf{monkey}\rangle, \langle\{2,1,1\}, \mathsf{q}\rangle, \langle\{1,2,1\}, \mathsf{rcd}\rangle, \langle\{1,1,2\}, \mathsf{esox}\rangle\}$$

$$L_{mixed} = \{\langle\{1,2,2,1\}, \mathsf{sghs}\rangle, \langle\{1,1,1,1\}, \mathsf{ppold}\rangle, \langle\{2,2,2,1\}, \mathsf{monkey}\rangle, \langle\{2,1,1,1\}, \mathsf{q}\rangle, \langle\{1,2,1,1\}, \mathsf{rcd}\rangle, \langle\{1,1,2,1\}, \mathsf{esox}\rangle, \langle\{1,2,2,2\}, \mathsf{adf}\rangle, \langle\{1,1,1,2\}, \mathsf{ace}\rangle, \langle\{2,2,2,2\}, \mathsf{bdf}\rangle, \langle\{2,1,1,2\}, \mathsf{bce}\rangle, \langle\{1,2,1,2\}, \mathsf{ade}\rangle, \langle\{1,1,2,2\}, \mathsf{acf}\rangle\}$$

Simplicity??

- explanatory simplicity relates to seeking the most economical explanation. For example, models developed with less parameters are preferred to those with more, and less ornate representations and process models are preferred to any others.
- cognitive simplicity relates to an assumption that cognitive structures fulfilling some function are consistently organised such that they are minimal in some sense.

scenario

- Simplicity as the driving force for the internal transformation of representations.
- Simplicity as the driving force behind induction from external linguistic stimuli.

Production and invention

- Production: If the unification of the set of meanings yields the meaning to be expressed, then the meaning can be expressed, and the resulting signal is formed by concatenating the symbols attached to each edge
- Invention is the process by which an agent produces an utterance containing a meaning which cannot be produced using the usual method.

Basic Theorems Used in Encoding

- Occam's Razor Principle
- Baye’s Theorem
- Rissanen's MDL principle
  - The MDL principle states that the best hypothesis for some observed data D is the one that minimises 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.

$$H_{MDL} = \min_{H \in \mathcal{H}} \{L_{C_1}(H) + L_{C_2}(D|H)\}$$

Basic Theorems Used in Encoding

- Two grammars for the same language might be equally complete and correct, and we wish our general theory to provide a criterion of choice between them. Of two theories that cover the same facts, it is usual to choose that one which is simpler (or more economical). ---Bach

Transducer

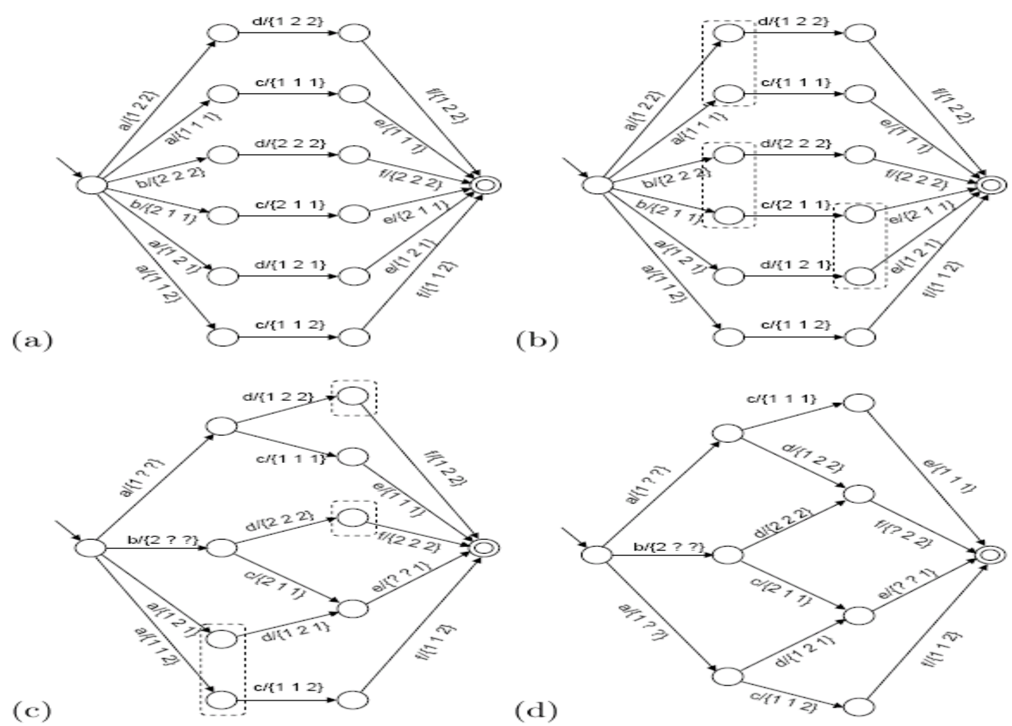
- An FSUT is a variation on the basic notion of a finite state transducer. A finite state transducer maps one regular language to another by attaching output symbols to each state transition within the transducer. An FSUT extends this mechanism by mapping meanings to signals,.

$$L_{comp} = \{\langle\{1,2,2\},\mathrm{adf}\rangle, \langle\{1,1,1\},\mathrm{ace}\rangle, \langle\{2,2,2\},\mathrm{bdf}\rangle, \langle\{2,1,1\},\mathrm{bce}\rangle, \langle\{1,2,1\},\mathrm{ade}\rangle, \langle\{1,1,2\},\mathrm{acf}\rangle\}$$

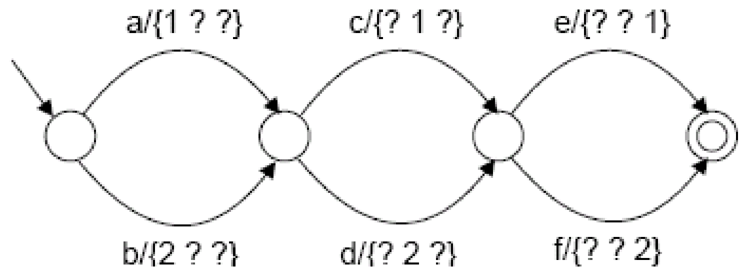
	value 1	value 2
feature 1	a	b
feature 2	c	d
feature 3	e	f

- State Merge: Two states q 1 and q 2 can be merged to form a new state if the transducer remains consistent. All edges that mention q 1 or q 2 now mention the new state.
- Edge merge: Two edges e 1 and e 2 can be merged if they share the same source and target states and accept the same symbol. The result of merging the two edges is a new edge with a new meaning label. Meanings are merged by finding the intersection of the two component meanings

$$L = \{\langle\{1,2,2\},\text{adf}\rangle, \langle\{1,1,1\},\text{ace}\rangle, \langle\{2,2,2\},\text{bdf}\rangle, \langle\{2,1,1\},\text{bce}\rangle, \langle\{1,2,1\},\text{ade}\rangle, \langle\{1,1,2\},\text{acf}\rangle\}$$



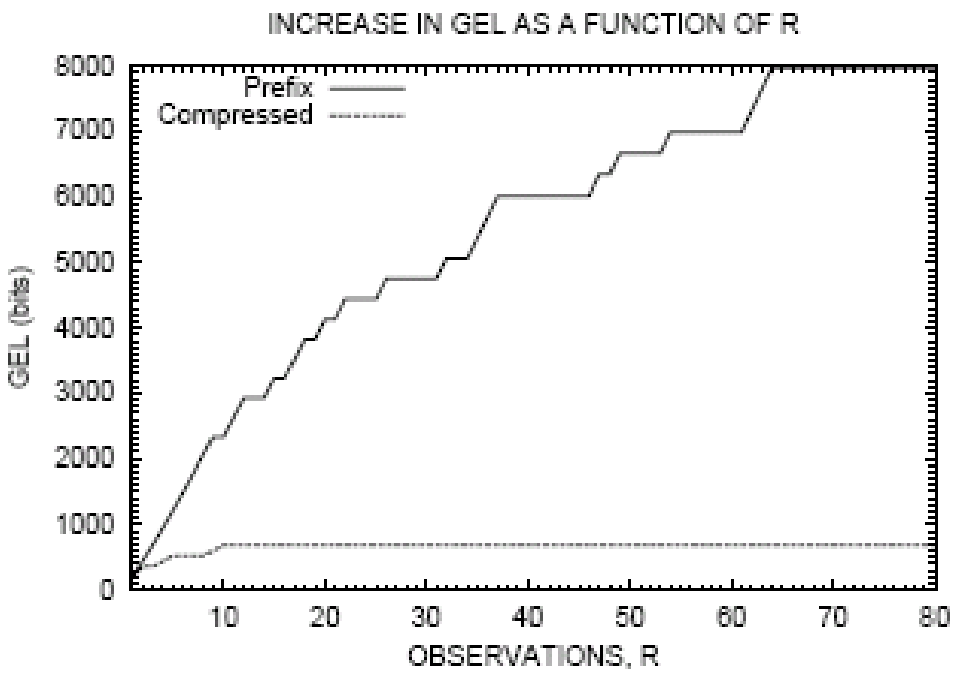
$$L_{comp} = \{\langle\{1,2,2\},\mathrm{adf}\rangle, \langle\{1,1,1\},\mathrm{ace}\rangle, \langle\{2,2,2\},\mathrm{bdf}\rangle, \langle\{2,1,1\},\mathrm{bce}\rangle, \langle\{1,2,1\},\mathrm{ade}\rangle, \langle\{1,1,2\},\mathrm{acf}\rangle\}$$

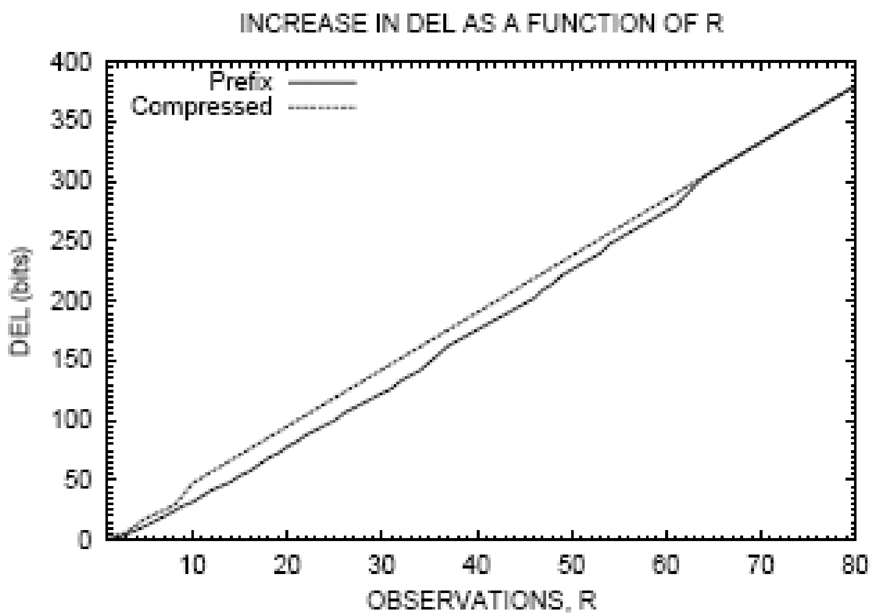


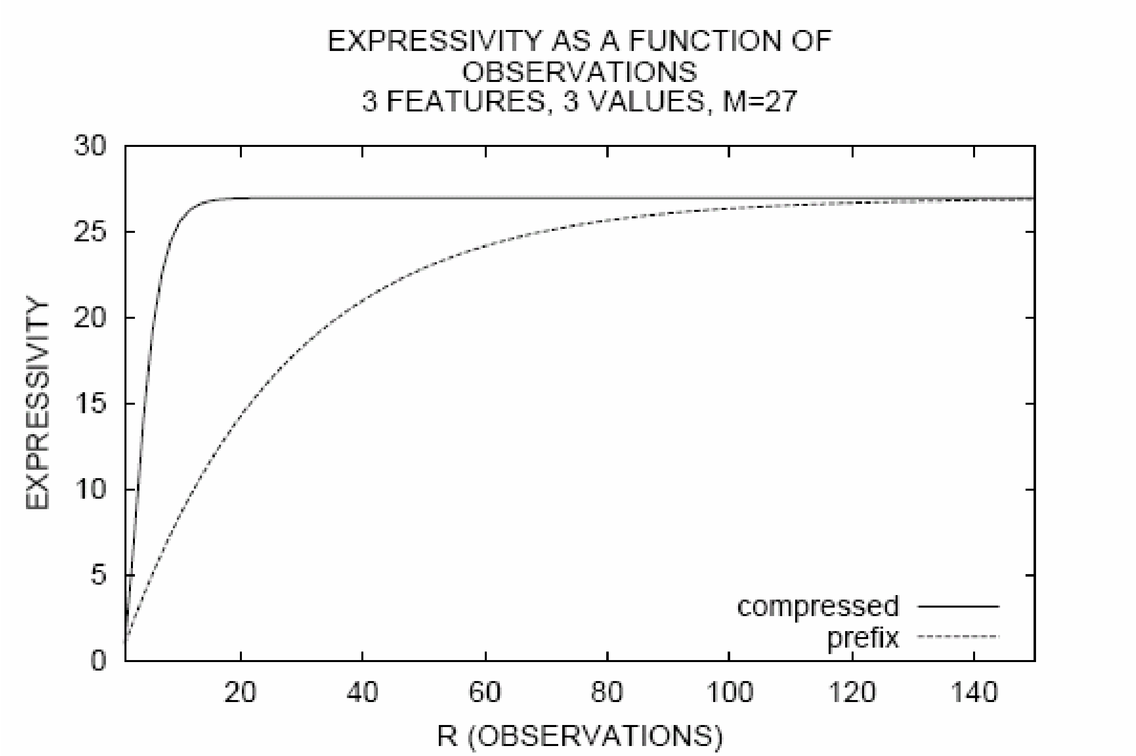
$$L^{+}_{comp} = \{\langle\{1,2,2\},\mathrm{adf}\rangle, \langle\{1,1,1\},\mathrm{ace}\rangle, \langle\{2,2,2\},\mathrm{bdf}\rangle, \langle\{2,1,1\},\mathrm{bce}\rangle, \langle\{1,2,1\},\mathrm{ade}\rangle, \langle\{1,1,2\},\mathrm{acf}\rangle, \langle\{2,1,2\},\mathrm{bcf}\rangle, \langle\{2,2,1\},\mathrm{bde}\rangle\}$$

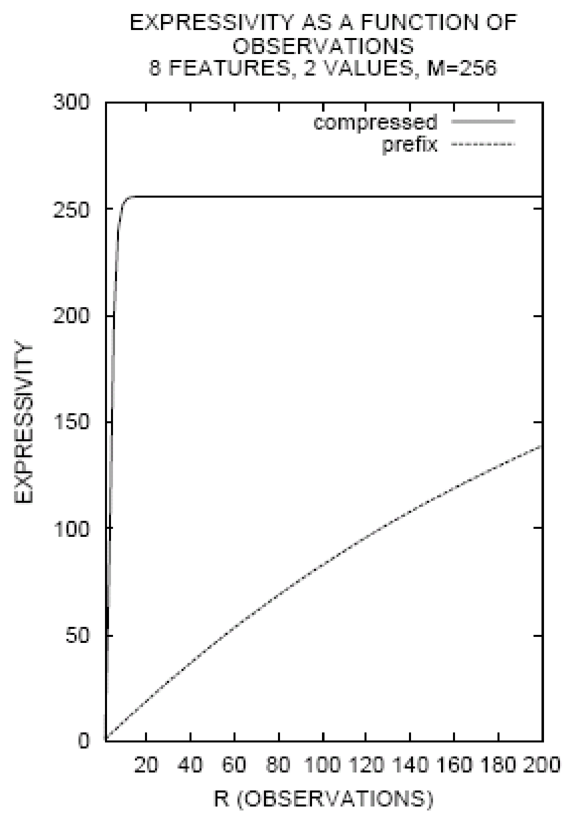
Assumption & Definition

- All meanings are equi-probable.
- Not all meanings are observed
- Expressivity: the number of meanings for which a signal can be constructed

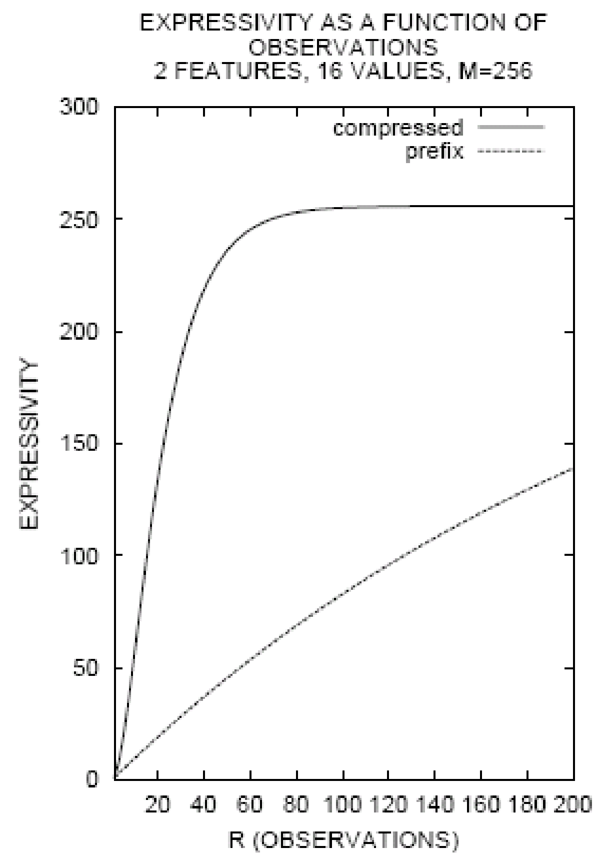

INCREASE IN GEL AS A FUNCTION OF R
Prefix
Compressed
GEL (bits)
0
1000
2000
3000
4000
5000
6000
7000
8000
10
20
30
40
50
60
70
80
OBSERVATIONS, R


INCREASE IN DEL AS A FUNCTION OF R
Prefix
Compressed
DEL (bits)
0
50
100
150
200
250
300
350
400
10
20
30
40
50
60
70
80
OBSERVATIONS, R

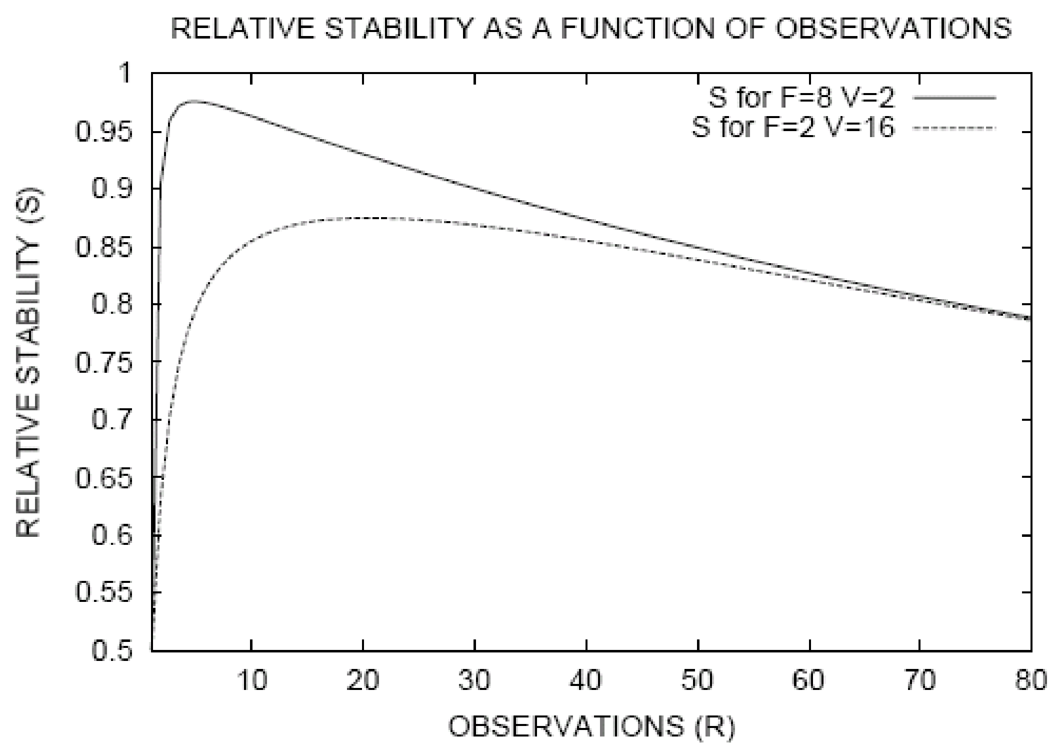

EXPRESSIVITY AS A FUNCTION OF
OBSERVATIONS
3 FEATURES, 3 VALUES, M=27
EXPRESSIVITY
R (OBSERVATIONS)
compressed
prefix
0
5
10
15
20
25
30
20
40
60
80
100
120
140


EXPRESSIVITY AS A FUNCTION OF
OBSERVATIONS
8 FEATURES, 2 VALUES, M=256
compressed
prefix
EXPRESSIVITY
R (OBSERVATIONS)
0
50
100
150
200
250
300
20 40 60 80 100 120 140 160 180 200

(a)


EXPRESSIVITY AS A FUNCTION OF
OBSERVATIONS
2 FEATURES, 16 VALUES, M=256
compressed
prefix
EXPRESSIVITY
R (OBSERVATIONS)
0
50
100
150
200
250
300
20 40 60 80 100 120 140 160 180 200

(b)


RELATIVE STABILITY AS A FUNCTION OF OBSERVATIONS
S for F=8 V=2
S for F=2 V=16
RELATIVE STABILITY (S)
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
10
20
30
40
50
60
70
80
OBSERVATIONS (R)

Key Points

- Explanation of Hallmark of languages
  - Compositionality
- Difference between animal and human language is justified
- Invention and production concepts are introduced.
- Transmission between generations is explained.

Drawbacks

- Which came first I-language or E-language
- How to deal with phrases and symbolic sentences
- can any language ever attain 100% compositionality??
- Interaction within a generation is not clear

Conclusion

- Whole analysis is based on
  – Sound Mathematical foundation
  – Logical reasoning
  – Good simulated result
- Still “does it depict real phenomena”??
  – Real life examples and scenario

Thanks!