Evolution of Language

Natalia Komarova

UC Irvine

Why work on language evolution?

- ... because the linguistic society of Paris officially banned any work on language evolution at a meeting in 1866.

Why work on language evolution?

- ... because one view is that language came as the by-product of a big brain.

Why work on language evolution?

- ... because one view is that language came as the by-product of a big brain.
- ... because another view is that language emerged suddenly, by one gigantic mutation.

What is remarkable about language ?

- It gives us unlimited expressibility.
- There are infinitely many grammatically correct sentences.
- It makes ‘infinite use of finite means.’ (W. von Humboldt)

What is remarkable about language ?

- You know about 60000 words.
- You learned about 1 new word per hour for 16 years.

What is remarkable about language ?

- Speech production is the most complicated mechanical motion we perform.
- Speech comprehension occurs at an impressive speed (up to 50 phonemes per second).

What is remarkable about language ?

- Talking is totally effortless.
- We can speak without thinking.

What is remarkable about language ?

- There are about 6000 languages.
- There is no simple language.

What is remarkable about language ?

- Language is the last of a series of major events that changed the rules of evolution.
  - 4 (?)   Origin of life (RNA,DNA,Proteins)
  - 3.5   Prokaryotes
  - 1.5   Eukaryotes
  - 0.6   Multicellular Organisms
  - 0.001   Language ⇨ new mode of evolution

(billion years ago)

What is new about our approach?

- Combine knowledge from the fields of linguistics, learning theory and evolutionary biology
- Use math a the main tool of description
- View language is a complex adaptive system

A complex adaptive system (following S. Levin)

- Consists of a number of different components;
- The components interact with each other with some degree of localization;
- An atonomous process exists that uses the outcomes of these interactions to select a subset of components for replication and/or enhancement.

In this talk

- Describe the micro-level (the level of individuals)
- Talk about the interactions of the components (the concept of individual learning)
- Include evolutionary forces and define group learning
- Discover new information about the bounds of complexity of human language

What is language?

- Mode of communication
- Crucial part of human behavior defining our social identity
- Uses combinatorial sequencing of small units into big ones

phonemes->syllables->words->phrases

->sentences

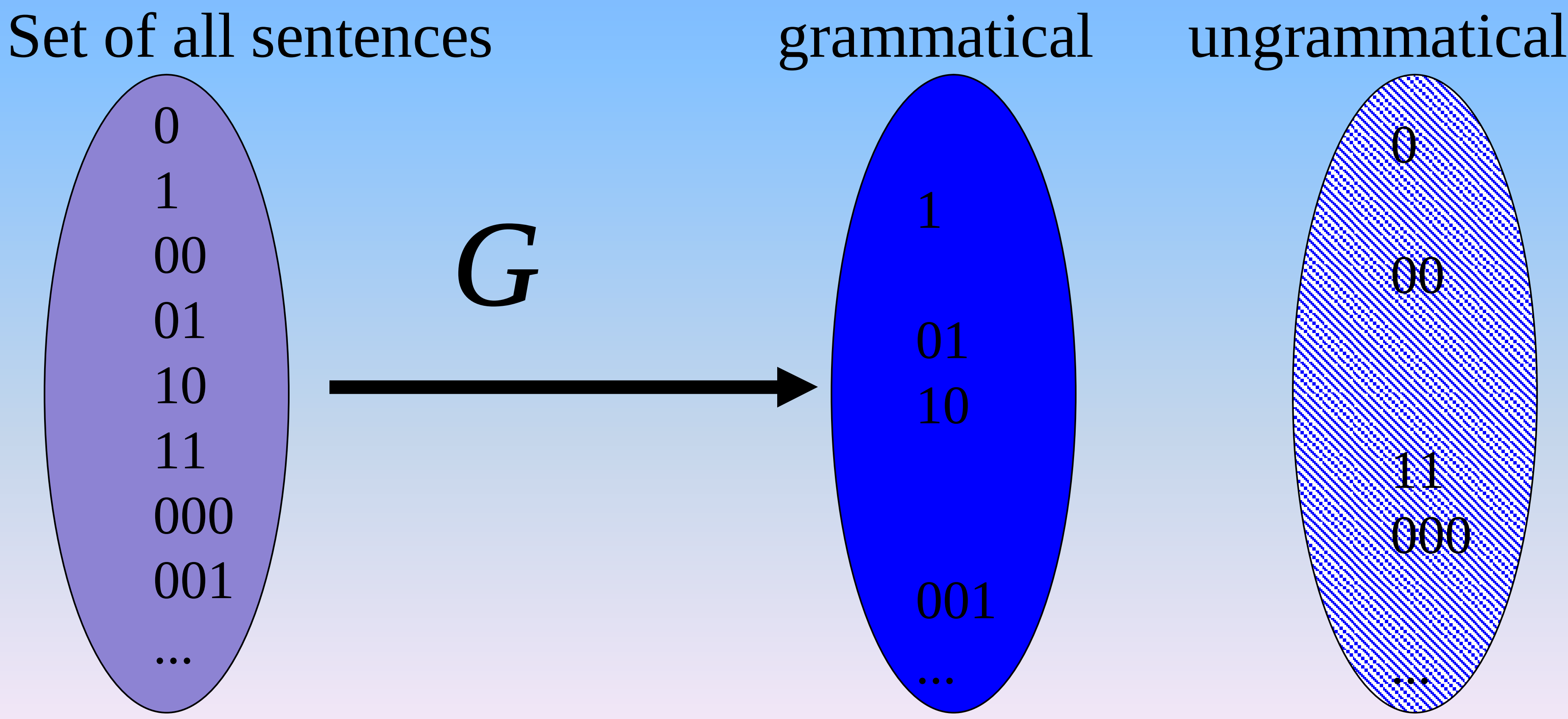
What is grammar?

- Grammar is the computational system of language
- Syntactic alphabet consists of a finite # of symbols, e.g. {0,1}
- Consider the set of all possible strings:

{0,1,00,01,10,11,000,001,...}

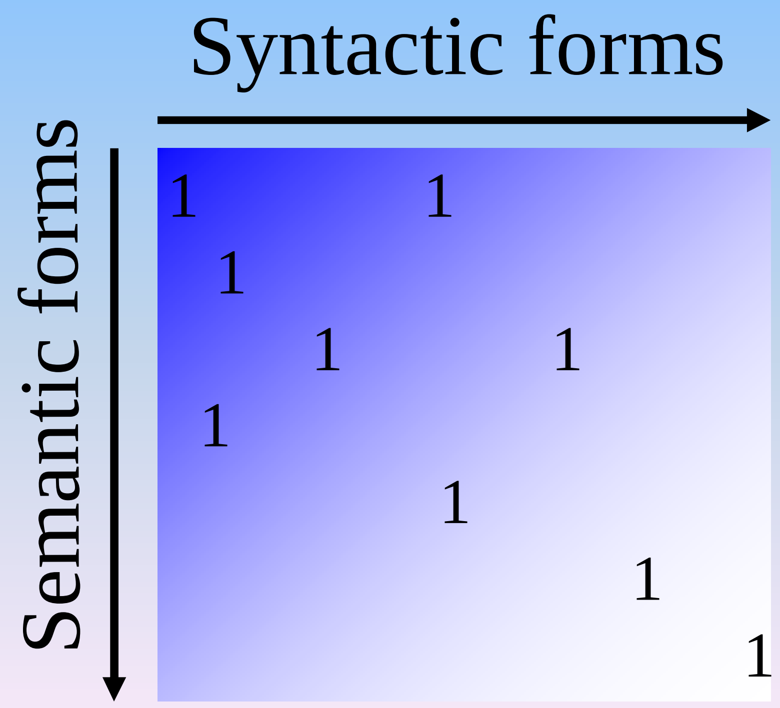
- A grammar is a rule system that specifies which strings are allowed

What is a grammar ?



Mapping between sound and meaning...

A grammar generates a mapping between syntactic forms and semantic forms



Grammar acquisition

- Children acquire the grammar of their native language by hearing grammatical sentences.
- This is only a small part of all possible sentences.
- This information does not uniquely determine the underlying grammatical rules.
- Nevertheless children reliably acquire the correct grammar.

Poverty of stimulus,
Paradox of language acquisition

Concept of individual learning

- What does it mean, to learn a rule?

Concept of individual learning

- A teacher generates a number of examples (applications of the rule), and the learner tries to figure out what the rule is.
- Learning="inductive inference"
- It is different from memorization because it allows the learner to generate new sentences
- Possibility to generalize...

Paradox of language acquisition

- I have a rule that generates certain integers, given an input.

Paradox of language acquisition

- I have a rule that generates certain integers, given an input.
- I will give several examples of correct application of the rule.

Paradox of language acquisition

- I have a rule that generates certain integers, given an input.
- I will give several examples of correct application of the rule.
- You guess the rule.

What is the rule?

- Input: 2 -> Output: 11

What is the rule?

- Input: 2 -> Output: 11
- Input: 3 -> Output: 111

What is the rule?

- Input: 2 -> Output: 11
- Input: 3 -> Output: 111
- Input: 4 -> Output: 1111

What is the rule?

- Input: 2 -> Output: 11
- Input: 3 -> Output: 111
- Input: 4 -> Output: 1111
- Input: 6 -> Output:

What is the rule?

- Input: 2 -> Output: 11
- Input: 3 -> Output: 111
- Input: 4 -> Output: 1111
- Input: 6 -> Output: 5811

What is the rule?

- Input: 2 -> Output: 11
- Input: 3 -> Output: 111
- Input: 4 -> Output: 1111
- Input: 6 -> Output: 5811
- The rule is:

Given x, calculate: $450\,x^2-2150\,x+2511$.

Paradox of language acquisition

- The rule ⟺ the grammar of the language
- The examples ⟺ the linguistic input (sentences) received by the child during the language acquisition stage
- Open-minded guessing does not work; the mathematical framework is given, e.g. by ***Gold's theorem****: no algorithm can learn a set of super-finite languages*

Paradox of language acquisition

- The fact that the linguistic input does not uniquely define the underlying grammar is referred to as "poverty of stimulus" (Wexler & Culicover 1980, Hornstein & Lightfoot 1981)
- The fact that children nonetheless manage to learn the grammar is termed "paradox of language acquisition" (Jackendoff 1997)

Universal grammar

- Children could not guess the correct grammar if they had no pre-formed, innate expectation.
- This innate expectation is **universal grammar**.

Noam Chomsky

Universal grammar

- We believe that the rule

$10^x+10^{x-1}+...+10^0$

is possible

- The rule

$450\,x^2-2150\,x+2511$

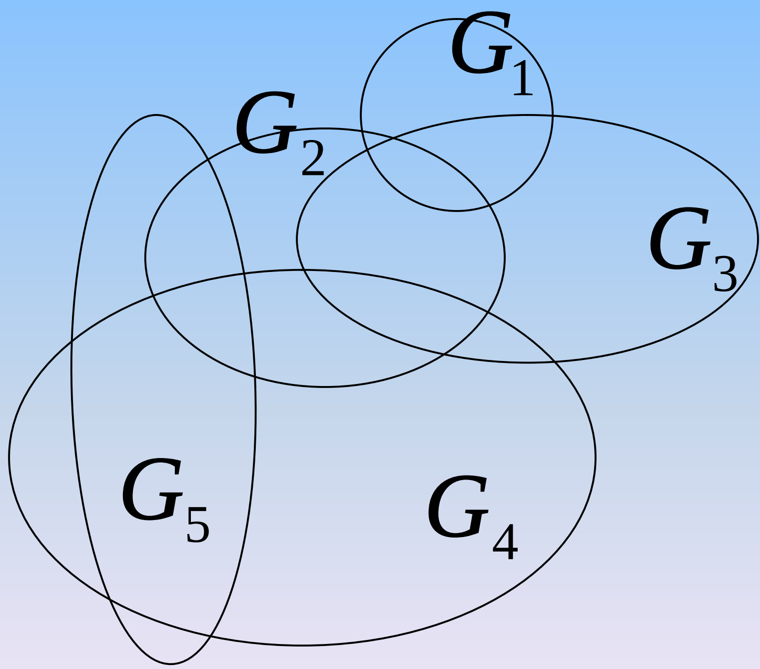
is impossible

Grammar acquisition

Universal Grammar

Candidate grammars

Learning
Mechanism:



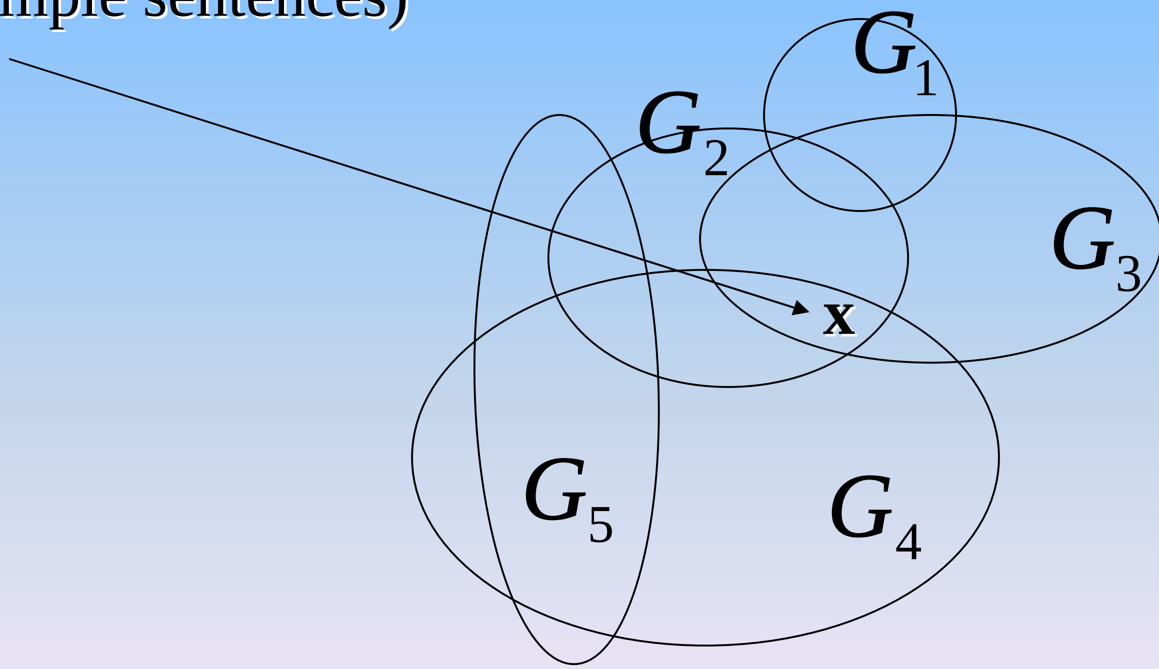
Grammar acquisition

Universal Grammar

Environmental input
(sample sentences)

Candidate grammars

Learning
Mechanism:
Which grammar
is it??



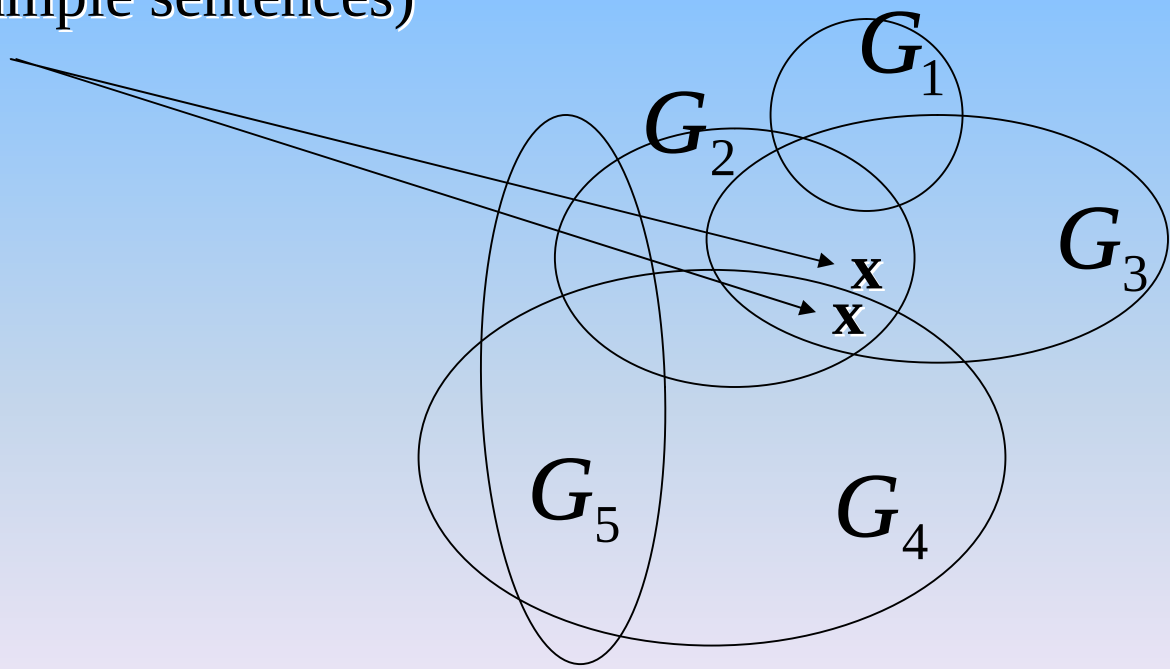
Grammar acquisition

Universal Grammar

Environmental input (sample sentences)

Candidate grammars

Learning Mechanism: Which grammar is it??



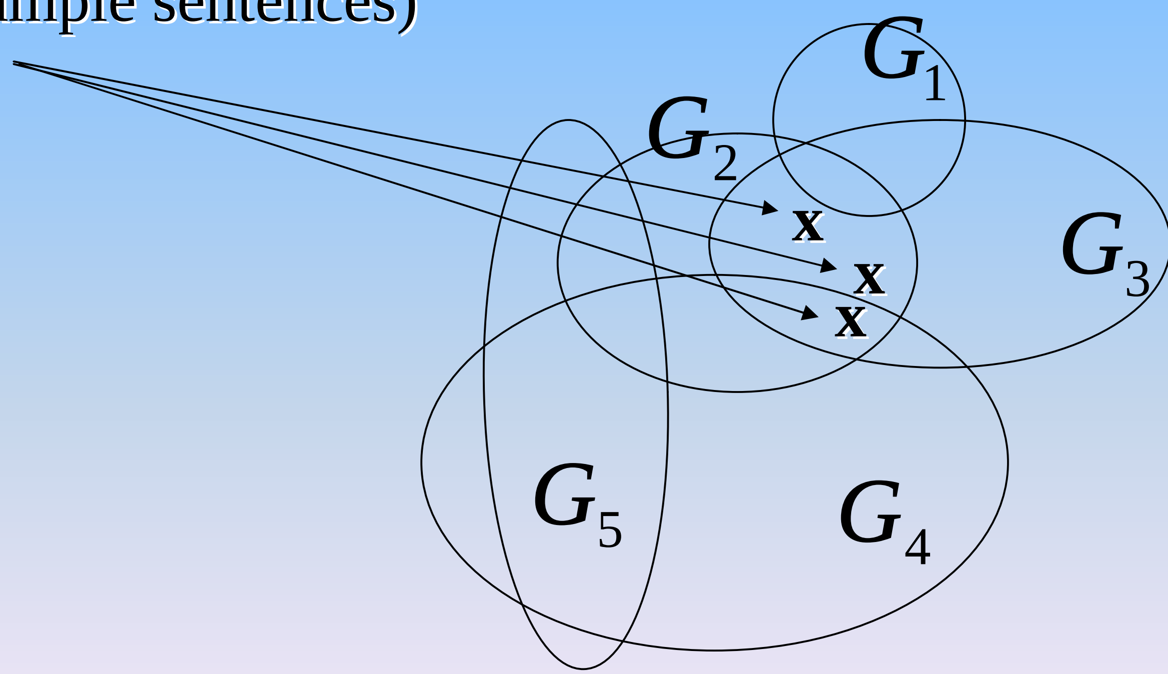
Grammar acquisition

Universal Grammar

Environmental input
(sample sentences)

Candidate grammars

Learning Mechanism:
Which grammar is it??



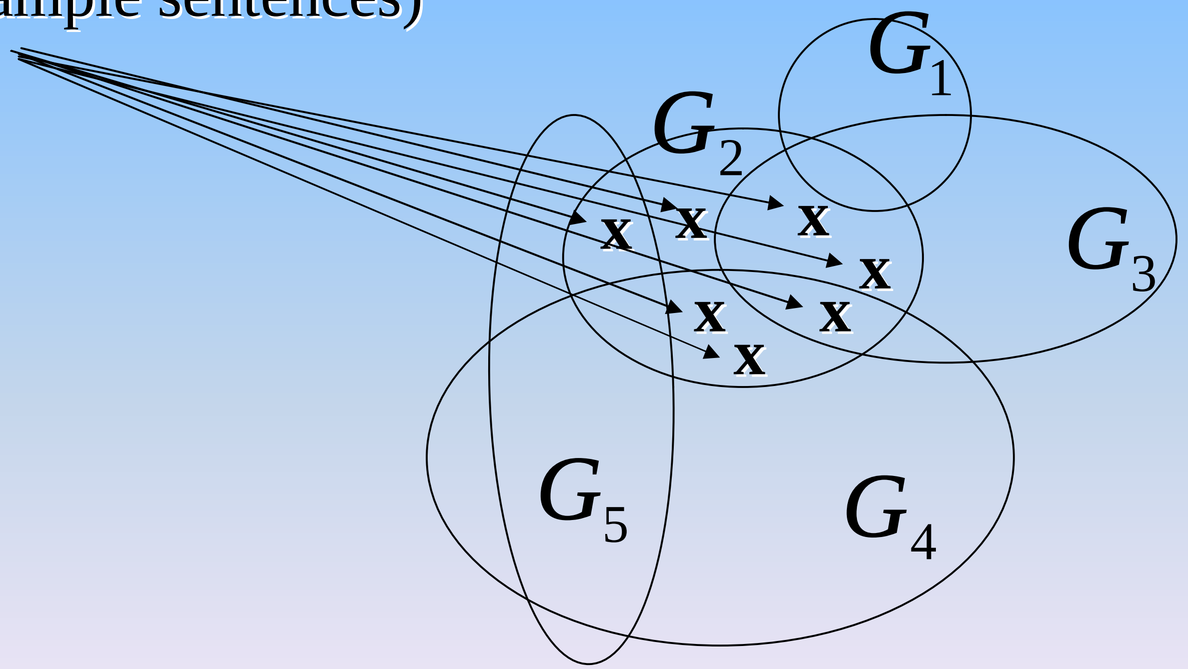
Grammar acquisition

Universal Grammar

Environmental input (sample sentences)

Candidate grammars

Learning Mechanism: Which grammar is it??



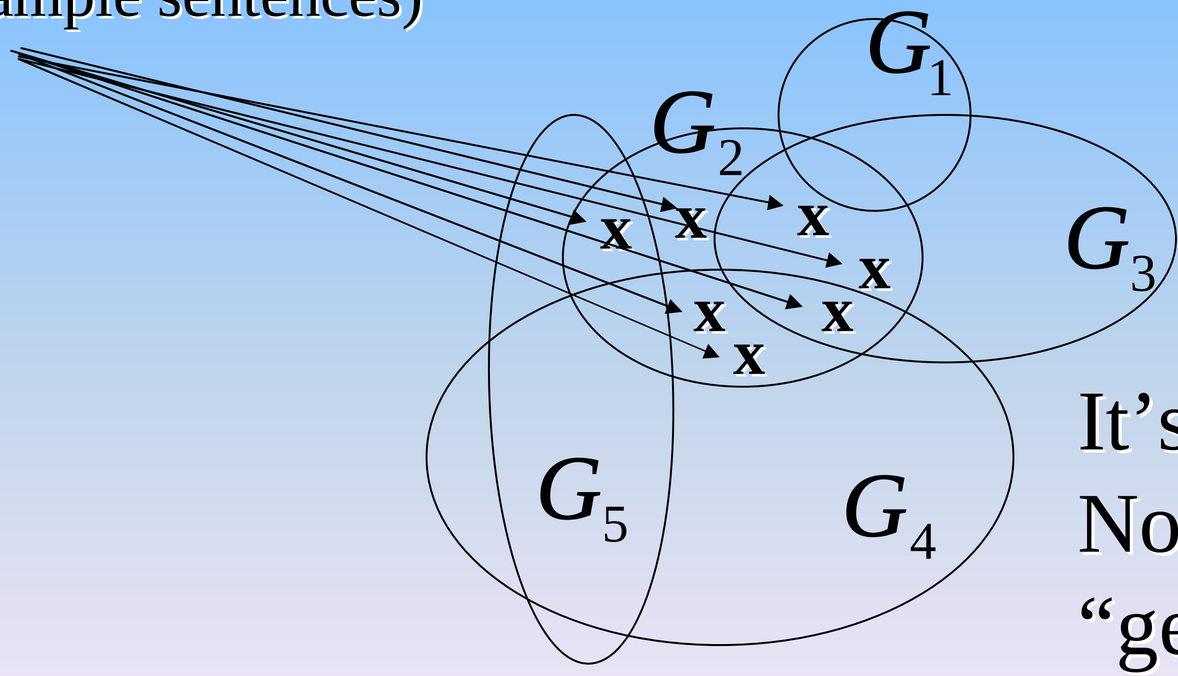
Grammar acquisition

Universal Grammar

Environmental input (sample sentences)

Candidate grammars

Learning Mechanism:
Which grammar Is it??



It's G_2!!!
Now the learner can "generalize" (utter new sentences).

Universal grammar

UG

Search space of candidate grammars

Learning mechanism to evaluate the environmental input

(memoryless learner, batch learner)

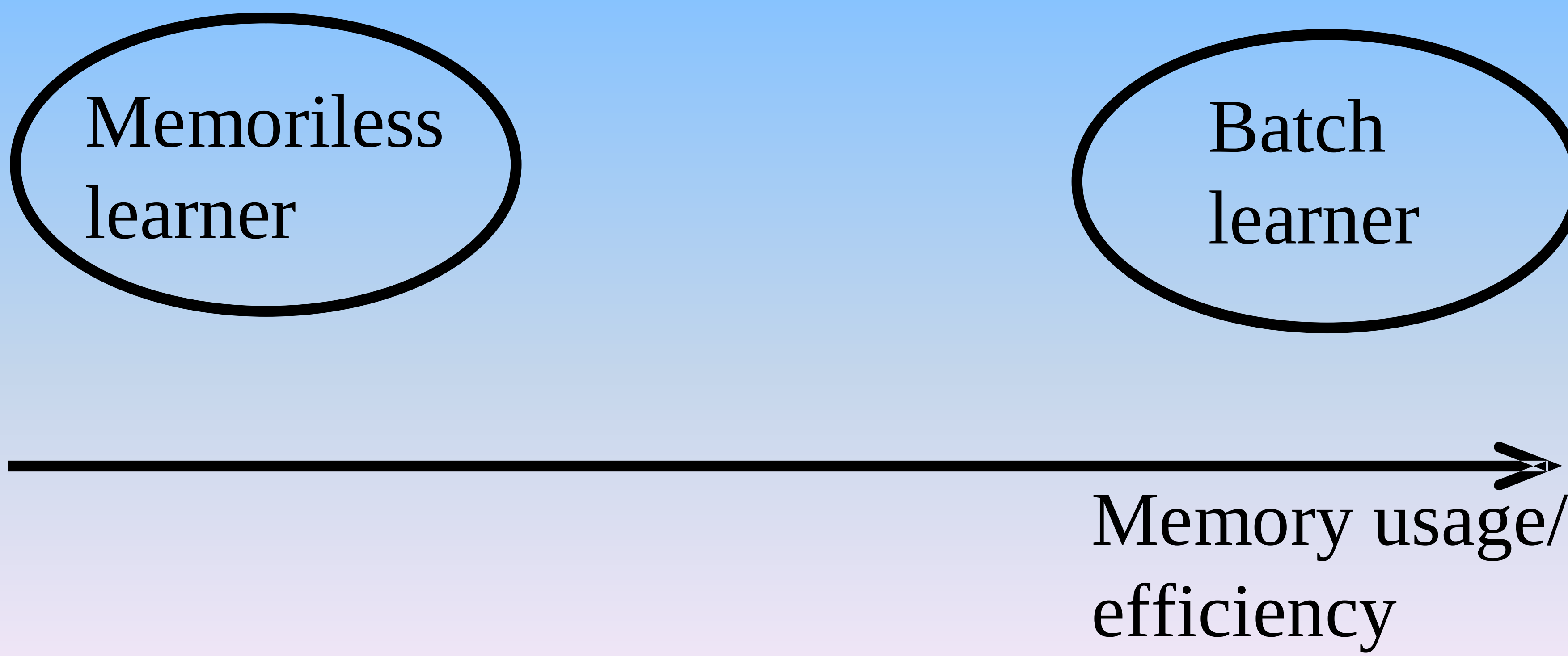
Memoryless learner

- Start with a randomly chosen grammar
- Stay with current grammar as long as sentences are compatible
- Change to a different grammar if a sentence is not compatible
- Stop after *N* sentences
- A very slow algorithm

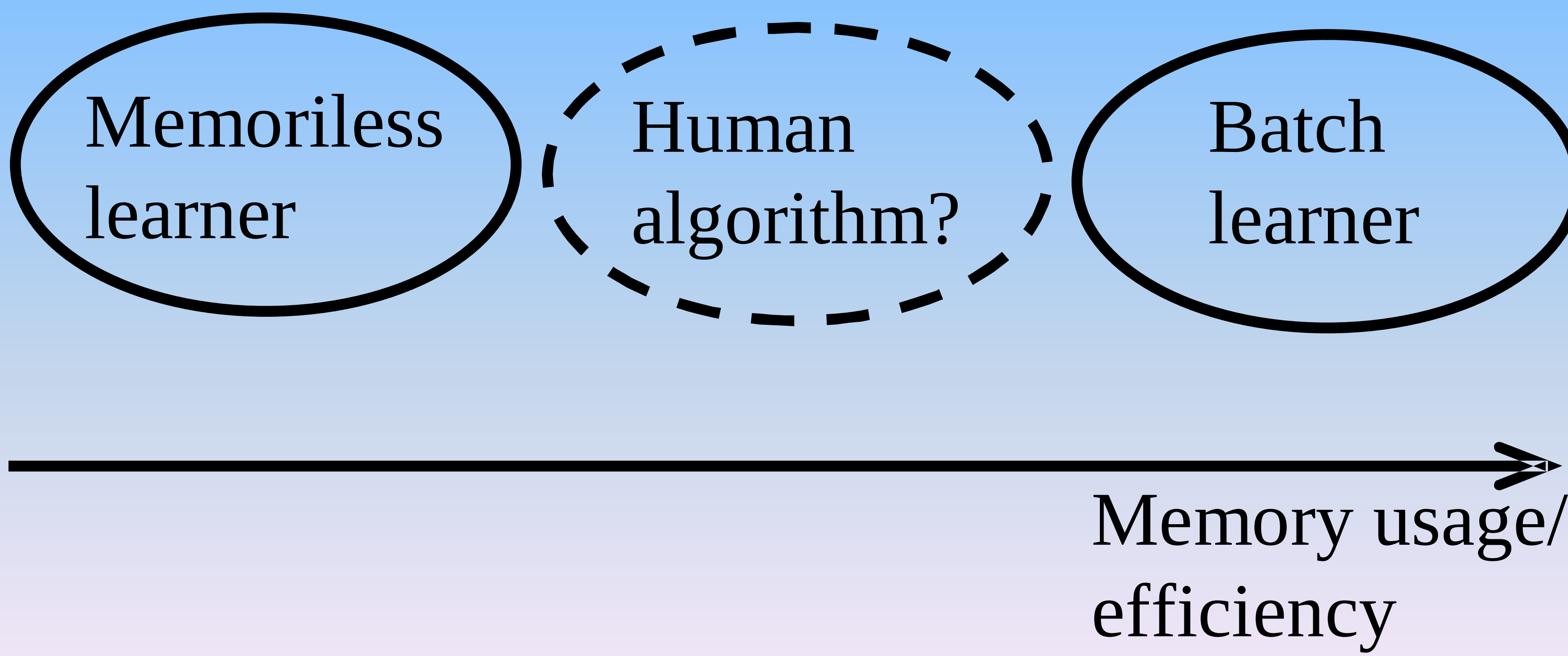
Batch learner

- Memorize all *N* sentences
- At the end, decide which grammar is most consistent with all these sentences
- A very efficient learner which requires an *infinite* memory

Human learning algorithm is unknown...



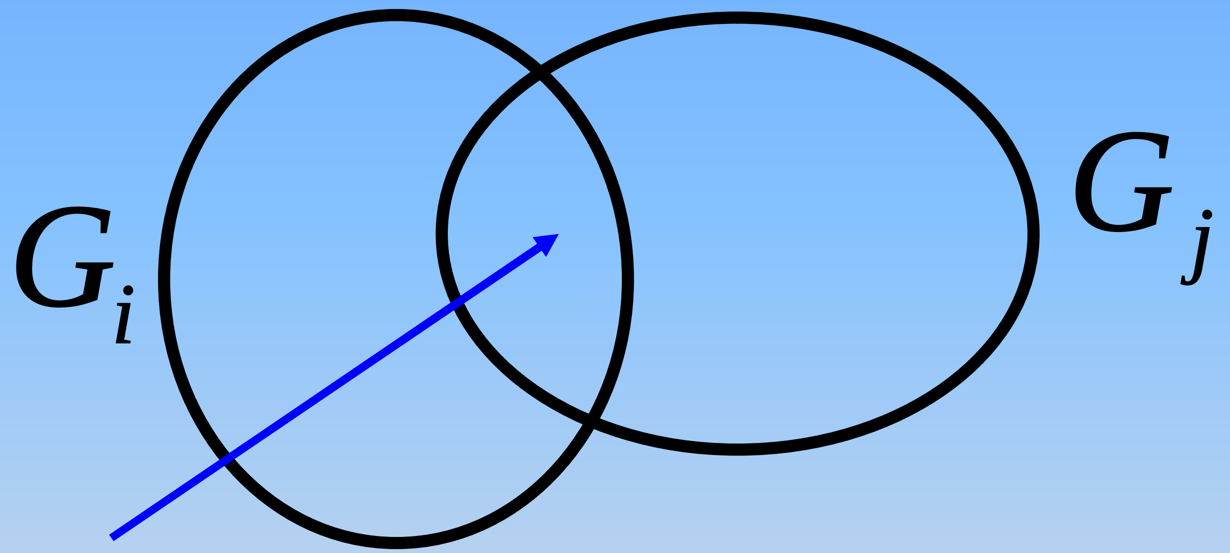
Human learning algorithm is unknown…



Convergence of individual learning algorithms

- Suppose we have n grammars, $G_1, \ldots, G_n$
- The teacher spits out N sample sentences
- Pairwise similarity of grammars is a random variable taken from a distribution
- How many sample sentences does, on average, a learner need to learn the correct grammar with confidence δ ?

Similarity of two grammars



$$s_{ij} = m_i(G_i \cap G_j)$$

probability that a speaker of G_i says a sentence that is compatible with G_j

Convergence of individual learning lagorithms

- The answer depends on n (the size of the "search space" of the UG) and on the distribution of pairwise similarities...
- For the memoriless learner,

$$N > C_1 |\log \delta|\ n \log n$$

- For the batch learner,

$$N > C_2\ F(\delta)\ n$$

How can this be useful?

- Individual learning: consider a teacher-learner pair; the time of convergence of learning algorithms can give bounds on the complexity of possible UG
- Population learning: consider a heterogeneous population of teachers and learners and its evolutionary dynamics: can coherence be reached and maintained?

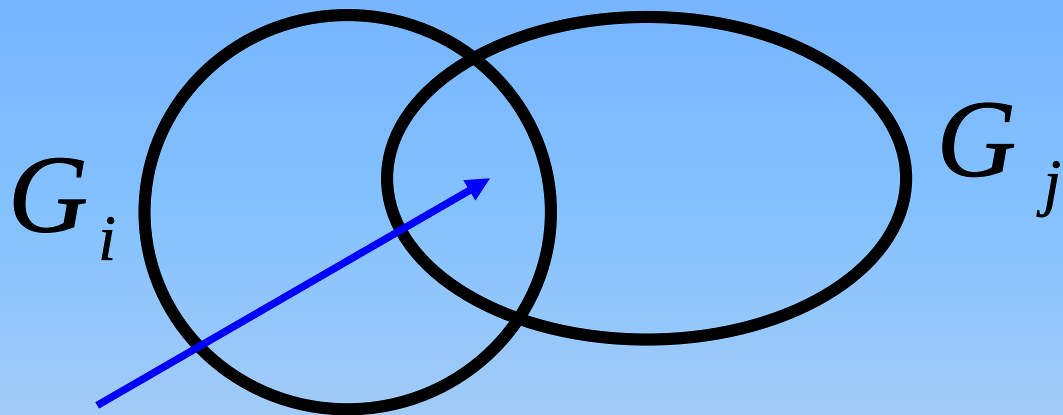
Population learning vs individual learning

- M individuals (M large)
- Individuals reproduce and die
- Children learn the language of their parents (individual learning!)
- Learning is not perfect (due to errors + innovations)
- Ability to communicate well is associated with biological fitness

Population learning

- Everybody starts off by speaking a (slightly) different language.
- Coherence is low.
- Run the evolutionary process.
- Under some circumstances, we can hope that the system might self-organize and converge to coherence...

Payoff for successful communication



$$s_{ij} = m_i(G_i \cap G_j)$$

$$F(G_i, G_j) = \frac{1}{2}(s_{ij} + s_{ji})$$

Language dynamical equation

$$\dot{x}_i = \sum_{j=1}^{n} x_j f_j(\vec{x}) Q_{ji} - f(\vec{x}) x_i$$

x_i ... frequency of G_i $\qquad \sum_{i=1}^{n} x_i = 1$

Fitness of G_i : $\quad f_i(\vec{x}) = \sum_{j=1}^{n} x_j F(G_i, G_j)$

Q_{ij}: probability that a learner will acquire G_j from a teacher with G_i

Language-dynamical equation

$$\dot{x}_i = \sum_{j=1}^{n} x_j \, f_j(\vec{x}) \, Q_{ji} - f(\vec{x}) x_i$$

$f(\vec{x}) = \sum_i x_i f_i(\vec{x})$... average fitness,

grammatical coherence

Language dynamical equation

$$\dot{x}_i = \sum_{j=1}^{n} x_j f_j(\vec{x})\, Q_{ji} - f(\vec{x})\, x_i$$

Selection Mutation

Nonlinear (3rd order), non-sparse system of n ordinary differential equations

Language dynamical equation = replicator-mutator equation

$$\dot{x}_i = \sum_{j=1}^{n} x_j f_j(\vec{x})\, Q_{ji} - f(\vec{x})\, x_i$$

- Autocatalitic reaction networks (f_i=reaction rates, expanded in terms of concentrations)
- Population genetics, with n alleles of a gene; $x_i x_j$ is the grequency of a gene pair

Other equations in evolutionary biology

Quasispecies equation

$$\dot{x}_i = \sum_{j=1}^{n} x_j f_j Q_{ji} - f(\vec{x}) x_i$$

Language equation

$$\dot{x}_i = \sum_{j=1}^{n} x_j f_j(\vec{x})\, Q_{ji} - f(\vec{x}) x_i$$

constant fitness

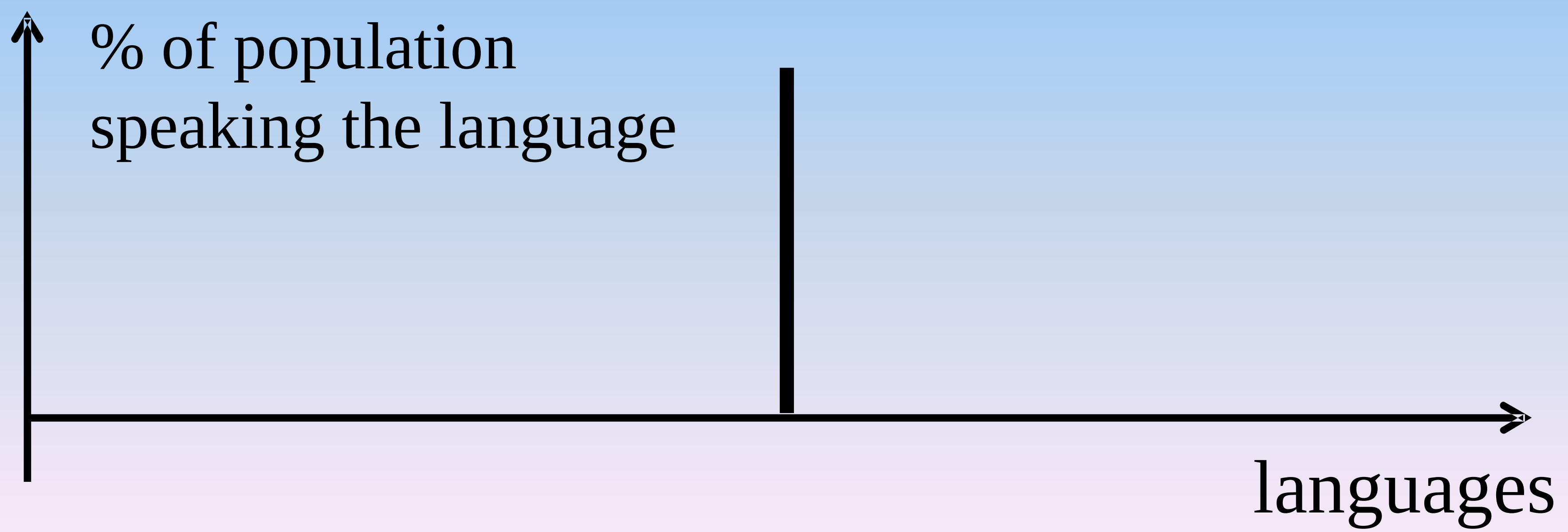
perfect learning

Replicator equation

$$\dot{x}_i = x_i [f_i(\vec{x}) - f(\vec{x})]$$

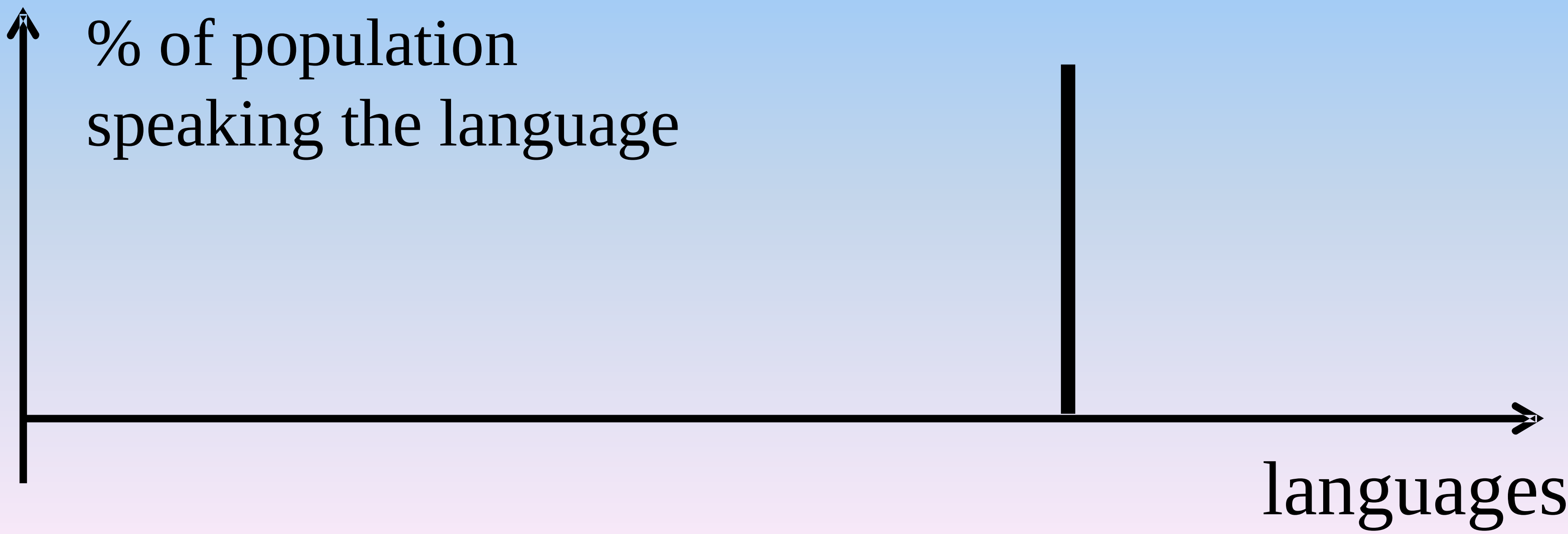
Dynamics of grammar acquisition

- Suppose there are no learning mistakes
- Then there are exactly n stable one-grammar “pure” solutions



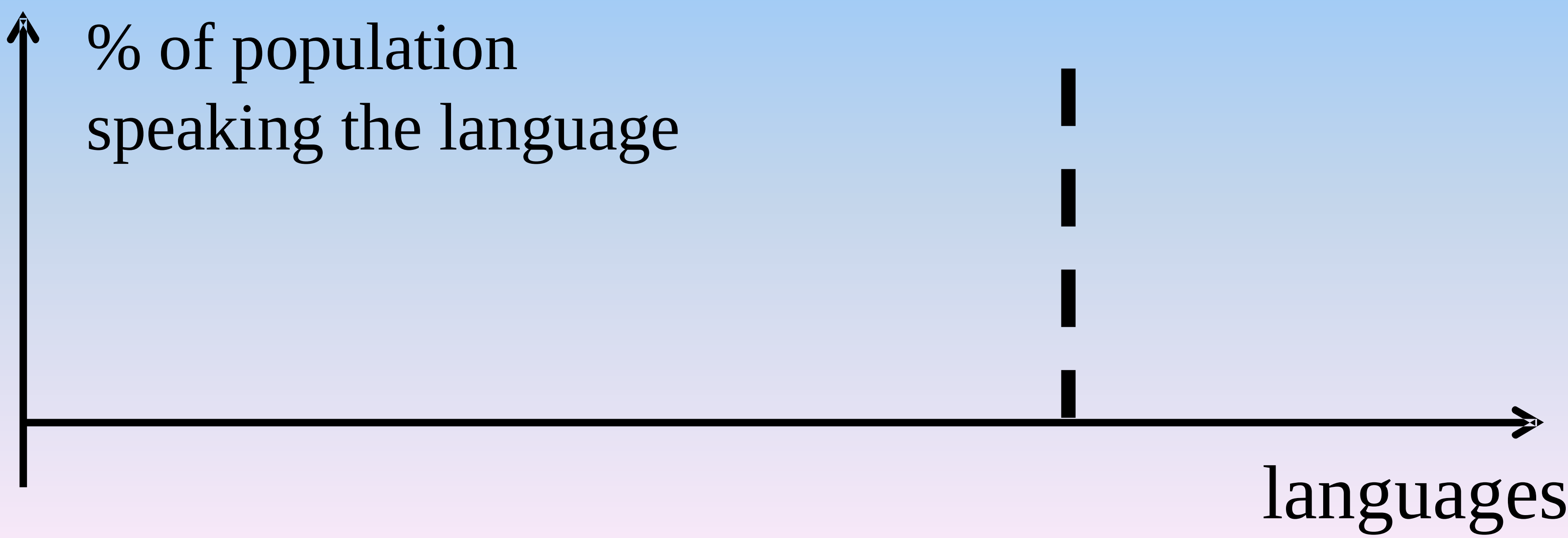
Dynamics of grammar acquisition

- Suppose there are no learning mistakes
- Then there are exactly n stable one-grammar “pure” solutions



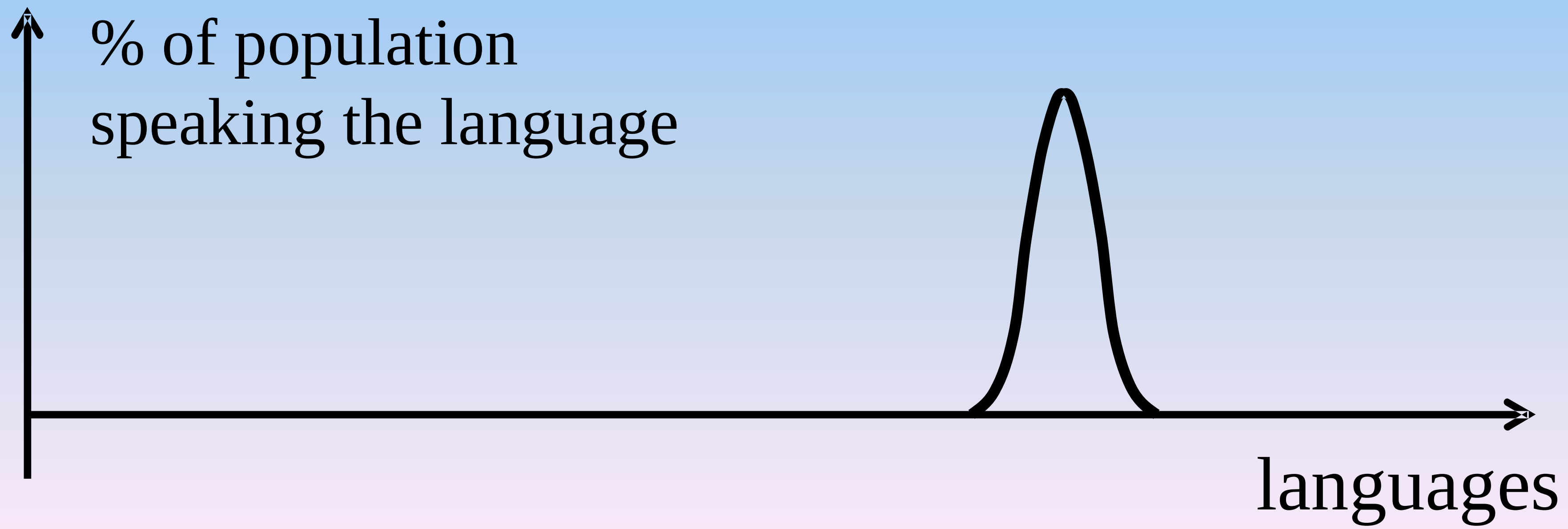
Dynamics of grammar acquisition

- Increase the chance of learning mistakes
- Then there are n stable “localized” one-grammar solutions



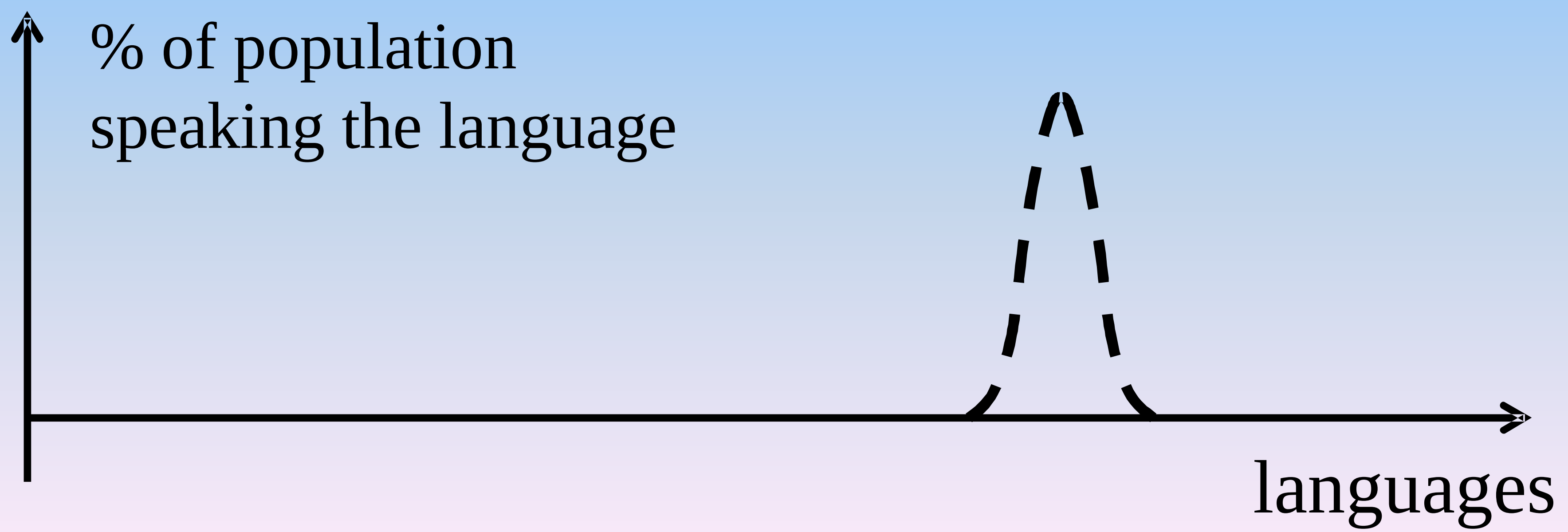
Dynamics of grammar acquisition

- Increase the chance of learning mistakes
- Then there are n stable "localized" one-grammar solutions



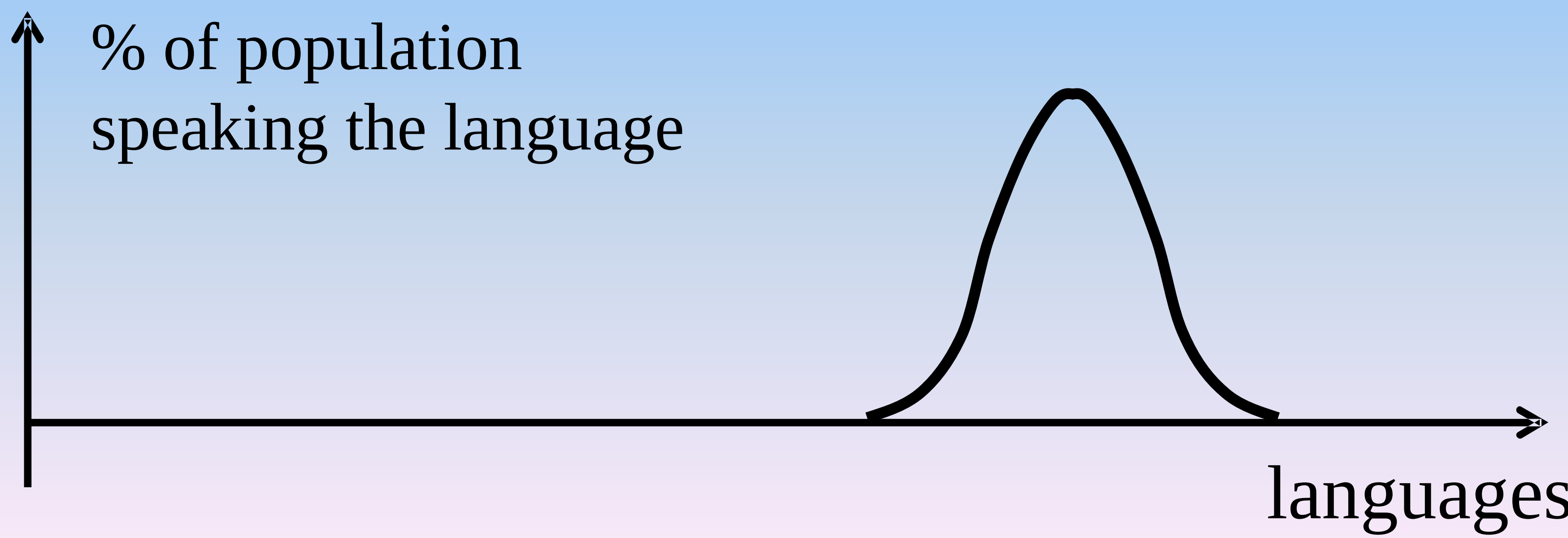
Dynamics of grammar acquisition

- Increase the chance of learning mistakes
- Then there are n stable “localized” one-grammar solutions



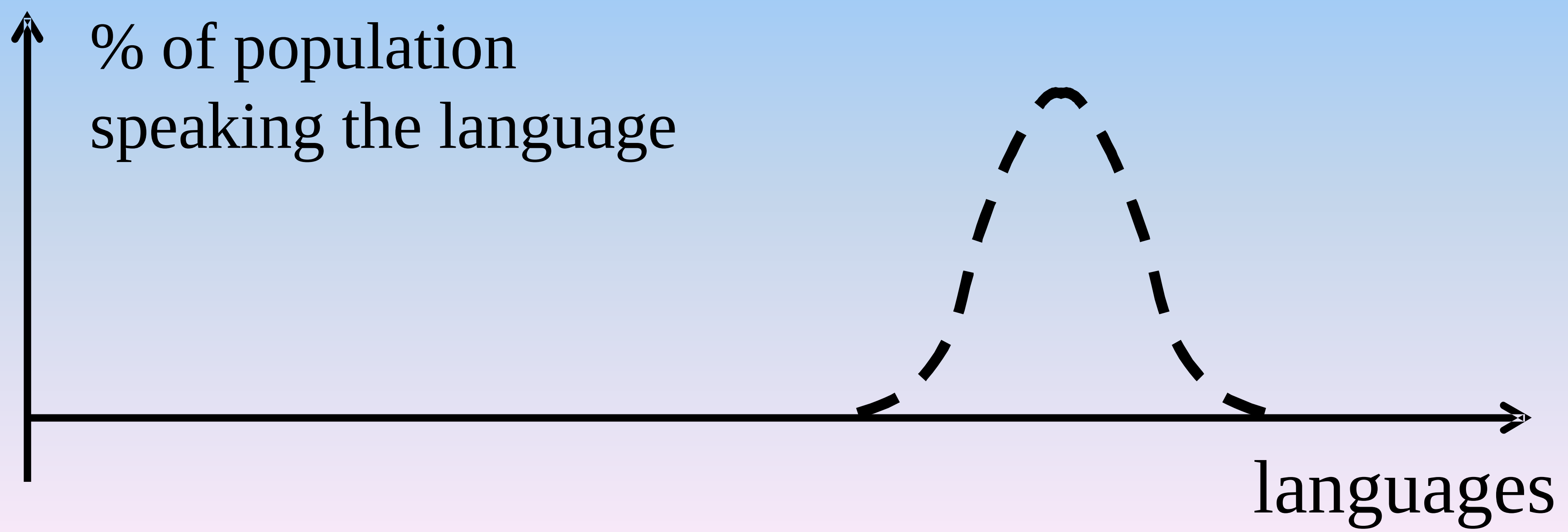
Dynamics of grammar acquisition

- Increase the chance of learning mistakes
- Then there are n stable “localized” one-grammar solutions



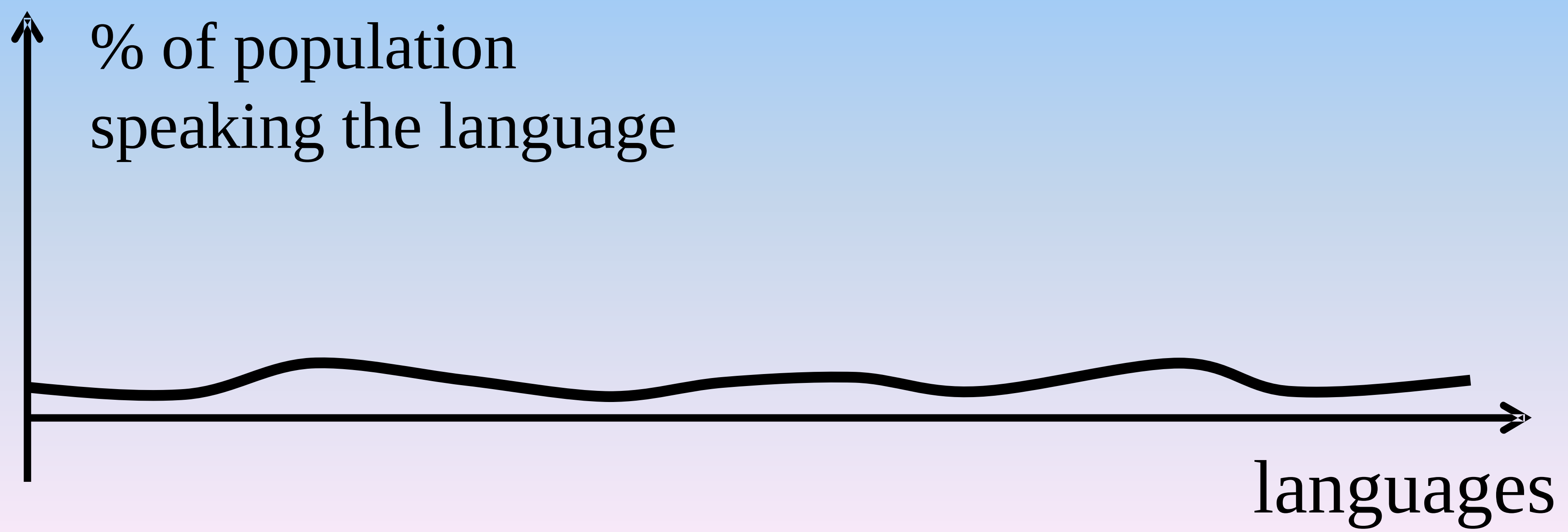
Dynamics of grammar acquisition

- Increase the chance of learning mistakes through a threshold value...
- Delocalization!



Dynamics of grammar acquisition

- Increase the chance of learning mistakes through a threshold value...
- Delocalization!



Dynamics of grammar acquisition: a bifurcation

- When learning accuracy is low, all grammars are present and coherence is low.

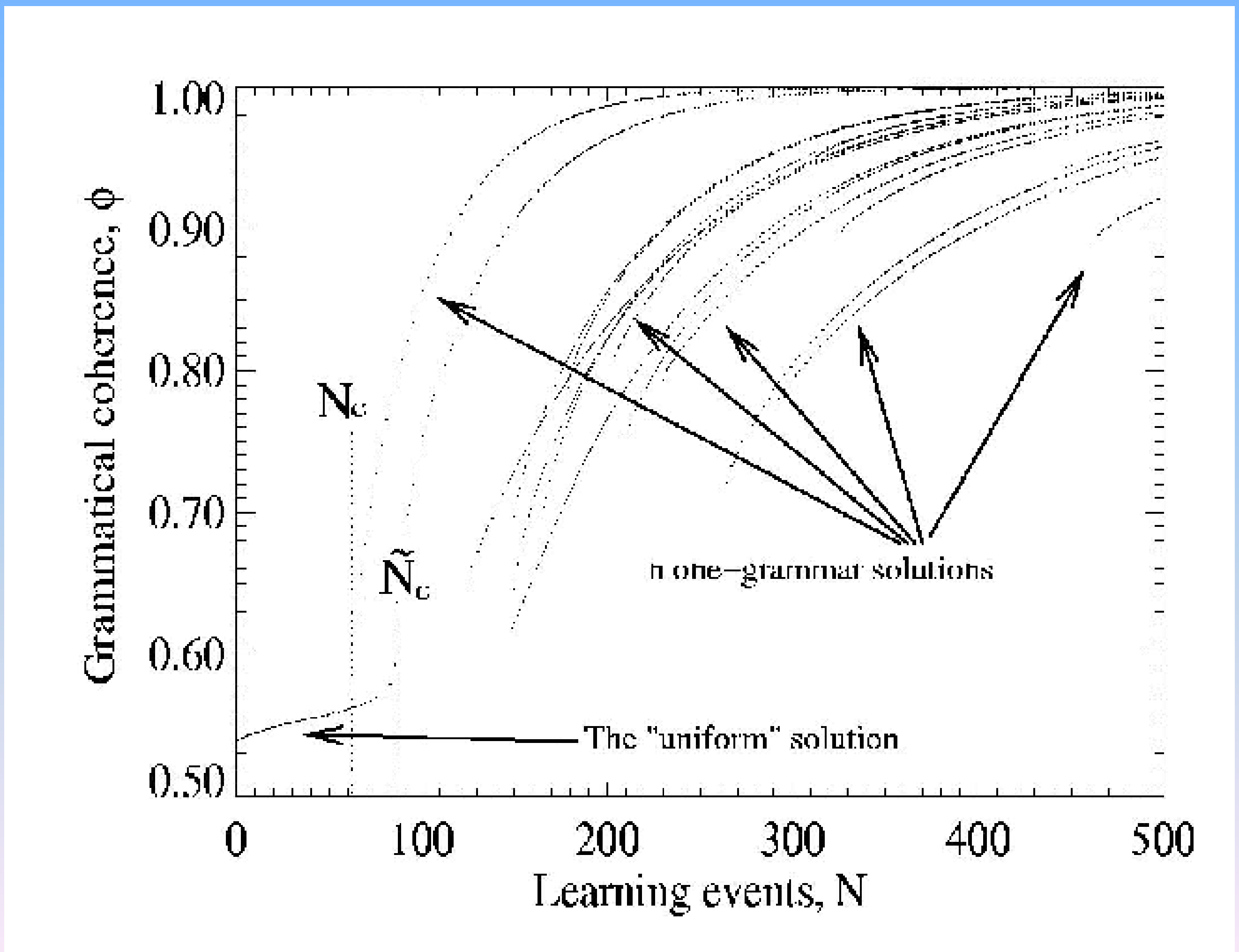
Dynamics of grammar acquisition: a bifurcation

- When learning accuracy is low, all grammars are present and coherence is low.
- For high learning accuracy, there are one-grammar equilibria, where most people speak exactly the same grammar. Coherence is high. There are many such equilibria → many possible languages.

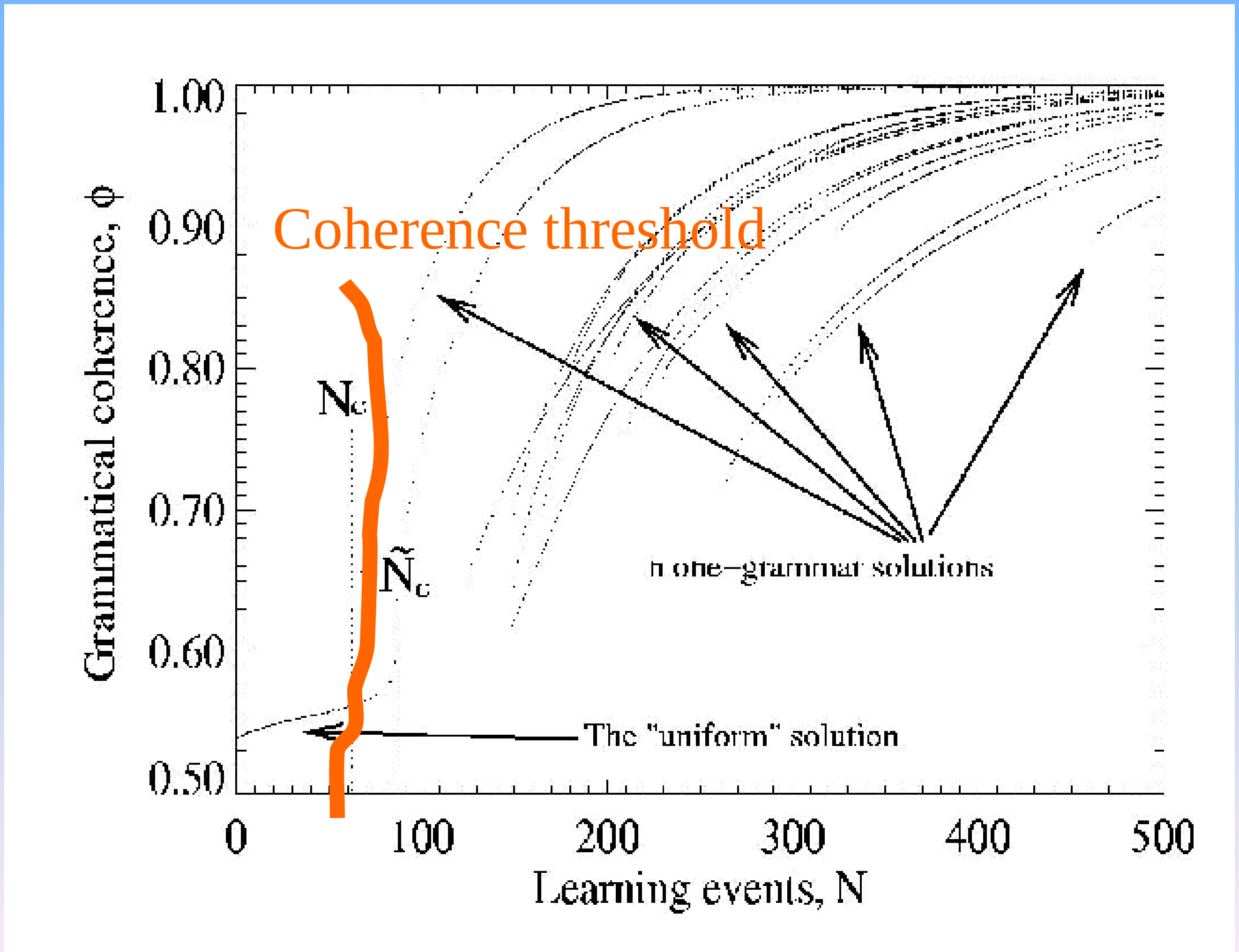
Number of sampling events and accuracy of learning

- Accuracy of learning increases with the number of sampling events, *N*, available to the child
- Low number of sampling events – low coherence in the population
- Large number of sampling events – existence of stable one grammar equilibria and high coherence

Coherence threshold



Coherence threshold



Universality property

- Coherence threshold does not depend on the size of Universal Grammar (if n is large).

Population learning: what are the questions?

- Individual learning: how many sample sentences does one need in order to learn the correct grammar with confidence δ?
- Population learning: how many sample sentence do children need in order for the population to be within Δ of perfect coherence?

Coherence threshold

If the accuracy of learning is high enough (or, if the number of sample sentences is large enough) then the population can develop and maintain a coherent grammatical system.

How is the coherence threshold related to the complexity of universal grammar?

Coherence threshold

- For a population of memoryless learners, we need

$$N > C_1 n \log n$$

sampling events

(n is the size of UG)

Coherence threshold

- For a population of memoryless learners, we need
$$N > C_1 n \log n$$
sampling events
(*n* is the size of UG)
- For batch learners we need
$$N > C_2 n$$

Coherence threshold

- For humans, the learning mechanism is unknown, but it lies somewhere between a memoryless learner and a batch learner.

  The minimum number of sampling events for humans is between

  $N = C_2 n$ and $N = C_1 n \log n$

Coherence threshold

- The minimum number of sampling events for humans is between

$$N = C_2 n \quad \text{and} \quad N = C_1 n \log n$$

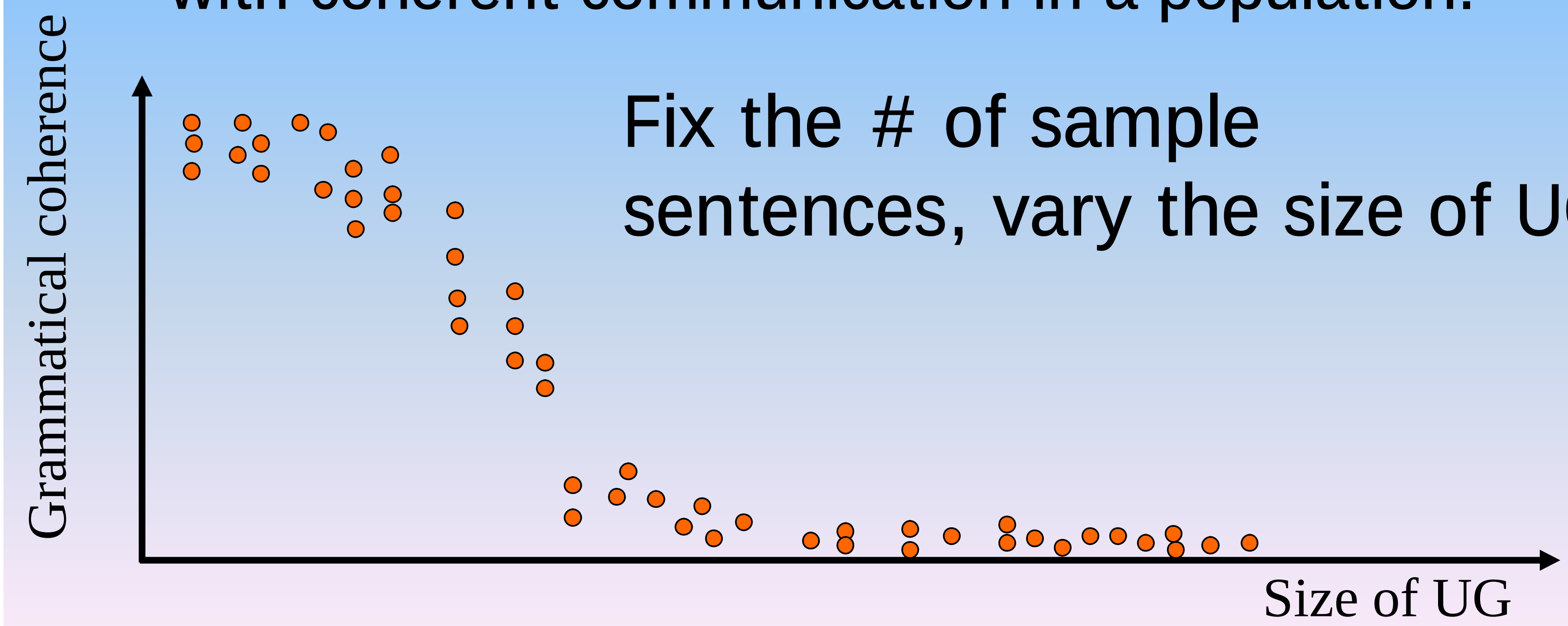
\# of sample sentences: defined by lifestyle

Complexity of Universal Grammar

The maximum complexity of the search space

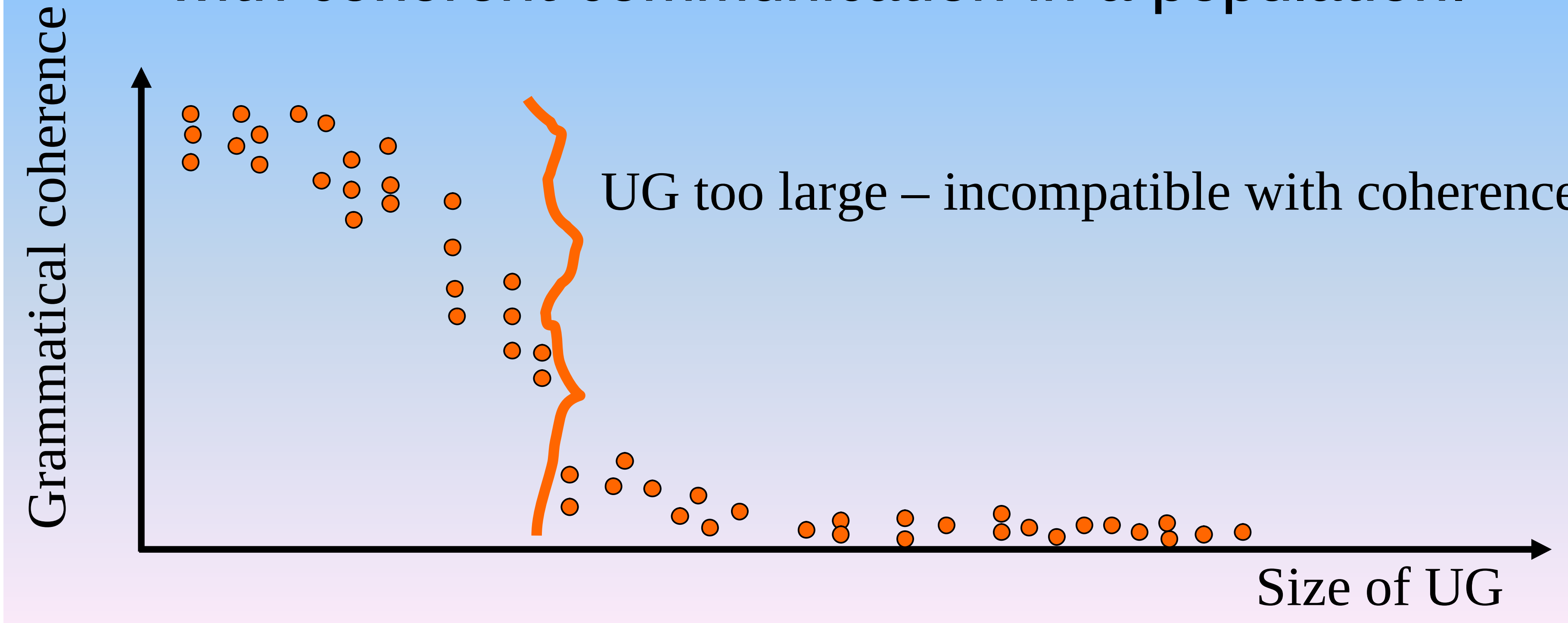
- We have an implicit condition for the maximum complexity of the search space that is compatible with coherent communication in a population.

Fix the # of sample sentences, vary the size of UG



The maximum complexity of the search space

- We have an implicit condition for the maximum complexity of the search space that is compatible with coherent communication in a population.



The maximum complexity of the search space

- Our conditions specify the maximum complexity of a universal grammar compatible with evolution.
- This is a necessary condition for UG to evolve by natural selection.

Conclusions

- Evolution and acquisition of grammar can be described using the formalism of Darwinian biology
- This involves the concept of population learning and the usage of a replicator-mutator equation

Conclusions

- Language is a self-organizing complex adaptive system, where a population of learners may reach and maintain coherence

Conclusions

- The level of coherence depends crucially on the learning accuracy
- There is a coherence threshold beyond which no coherence is possible
- If learning accuracy is sufficient, a coherent language appears as a stable equilibrium. Many such equilibria are possible

Conclusions

- Learning accuracy, the complexity of UG and the amount of linguistic input available to children must satisfy certain conditions for UG to have evolved

Collaborators

- Martin Nowak (Harvard)
- Partha Niyogi (Chicago)
- Igor Rivin (Temple)
- Ray Mendoza (UCI)
- Dirk Groeneveld (UCI)

What is wrong with Gold's theory?

- The target language has to be identified exactly
- Only positive examples
- Access to an arbitrary large # of examples
- No considerations of computational complexity

What is wrong with Gold's theory?

- The target language has to be identified exactly (restrictive)
- Only positive examples (restrictive)
- Access to an arbitrary large # of examples (unrestrictive)
- No considerations of computational complexity (unrestrictive)

Other theories

- The target language has to be identified exactly (restrictive)
- Only positive examples (restrictive)
- Access to an arbitrary large # of examples

(unrestrictive)

Statistical learning theory

- No considerations of computational complexity (unrestrictive)

Other theories

- The target language has to be identified exactly (restrictive)
- Only positive examples (restrictive)
- Access to an arbitrary large # of examples (unrestrictive)
- ~~No considerations of computational complexity (unrestrictive)~~

Valiant's theory