



# 语言产生的计算仿真研究

## Simulation Perspective on Language Emergence

---

**龚涛 Tao Gong**

**DSP与语音技术实验室 DSP and Speech Technology Laboratory,**  
**香港中文大学电子工程系 Dept. of EE, CUHK**

2005年5月13日 May.13, 2005;



## 摘要 outline

- 计算仿真工具—多个体系统  
Computational Simulation -- Multi-agent System
- 模型设计及模型示例—词汇-语法共同演化  
Design of model, and Example – Lexicon-syntax  
coevolution
- 讨论与总结  
Discussion and Conclusion



# 语言学研究方法(Research Method)归纳

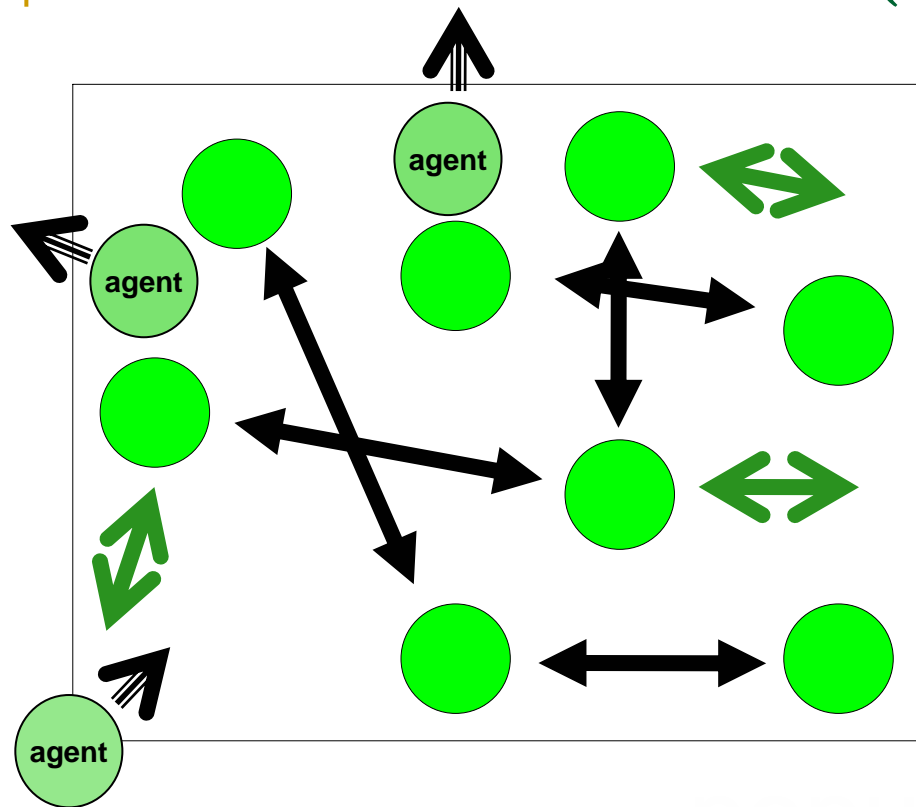
## ■ 传统方法(Traditional):

- 考古学(Archaeology), 人类学(Anthropology)发现和生物学(Biology)试验;
- 语言数据采集和语言库(Language Database)的建立;

## ■ 新方法(New):

- 计算机模拟(Computational Simulation)
  - 多个体系统 (Multi-agent system);
  - 人工神经网络 (Artificial Neural Network);
- 用计算机模拟研究语言演化问题的小组  
(major groups of computational simulation on linguistic problems)
  - Language Evolution and Computation Research Unit (LEC),  
Edinburgh University (Jim Hurford and Simon Kirby);
  - Artificial Intelligence Laboratory, Vrije Universiteit Brussel (Luc Steels);
  - Santa Fe Institute (SFI) (John Holland);
  - Language Engineering Laboratory (LEL), CUHK (William Shi-yuan Wang 王士元);

# 计算机仿真—多个体系统(Multi-agent System)



## 个体(Agent) (Steels 1999):

- 1) 具有独立性(Autonomous entity);
- 2) 具备某些能力(abilities);

## 个体的能力(Abilities):

- 1) 记忆体 (Memory): 如, 用规则记录语言;
- 2) 活动 (Activities): 如, 与其他个体交流 (Interaction), 被新个体替换 (Replacement);
- 3) 每个个体具有独特性 (Heterogeneity);

## 目标(Objectives): 测试通过交流后的整体特性:

- 1) 共同使用相似的规则记录语言(sharing of common rules);
- 2) 全局的结构(global structure);
- 3) 个体与所处环境的关系(human-environment relation);

agent



## 此类模型的特点 (Features of multi-agent system)

- 适宜研究个体行为(individual activities)所导致的集体表现(collective behavior);
- 适宜研究历时(diachronic)表现和系统发生(phylogenesis);
- 适宜研究群体间的通过个体所带来的影响(cross-group influences);

## 与语言产生问题的相似性 (Similarity to Language Emergence)

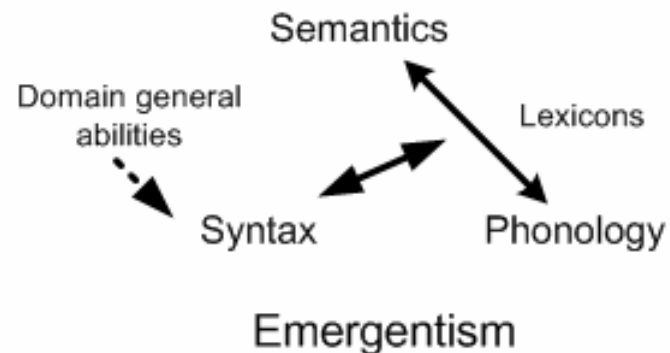
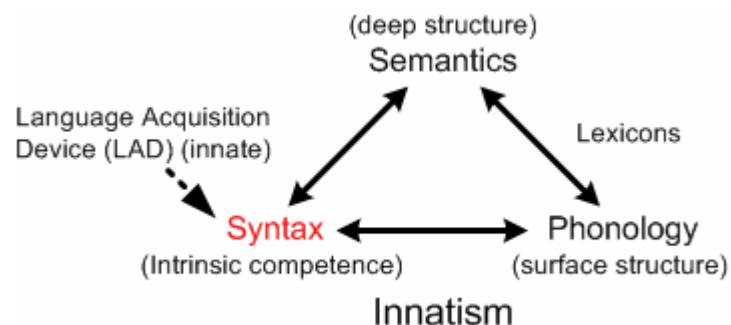
- 由下至上(bottom-up approach): 研究个体语言(language on individual level)对整体环境(environment)和群体语言(language on community level)之影响;
- 研究语言共有特性(linguistic features)的历时系统产生(diachronic, phylogenetic emergence)和其共时个体习得(synchronic, ontogenetic acquisition);
- 研究交流(communication)和社会结构(social structure)对语言的影响;
- 研究演化产生论(Emergentism)和天生观点(Innatism)对语言发展的解释;



# 语言产生理论背景—天生论和演化产生论

## Theoretical Background on Language Emergence -- Innatism vs. Emergentism

- **天生论 (Innatism)** (Chomsky 1995; Jackendoff 1997)
  - 语法是人类语言特有现象，是语义 (Semantics)和音韵(Phonology)的内在联系。语法(Syntax)由人类特有的语言习得装置 (Language Acquisition Device)决定。
- **演化产生论 (Emergentism)** (Knight et al. 2000):
  - 语法只是一种联系语音语义的机制
  - 语法是由一些人类和动物普遍具备的能力 (Domain-general abilities)演化而来的;



### 我们的研究重点 (Our approach):

采用多个体模型(multi-agent models)和自下而上的策略(bottom-up approaches)来探讨语言产生所需的最小限度的天生能力(minimal sets of innate capabilities)和基于此条件下的语言产生的动态特性(dynamics of such emergence).



## 模型设计及模型示例—词汇-语法共同演化

### Design of model, and Example – **Lexicon-syntax coevolution**

#### ■ 目的(objective)

- 语言产生过程：由整合信号系统(holistic signaling system)向合成语言(compositional language)的转变；
- 简单语法(主导词序(dominant/canonical word order))可否从某些普遍能力中(如排序能力(sequencing ability))转化而来；
- 词汇产生和主导词序收敛的具体过程；  
(emergence of lexicons and convergence of dominant word order)
- 基于交流的简单社会结构的产生  
(emergence of simple social structure based on communication);



## 模型设计及模型示例—词汇-语法共同演化

### Design of model, and Example – **Lexicon-syntax coevolution**

#### ■ 模型主要成分(**components**)

- **如何抽象语言**：规则系统(rule-based system);
- **个体及其行为**：
  - 规则的存储(memory);
  - 规则的获得(rule acquisition);
  - 语言交流和规则竞争(communication and rule competition);
- **演化过程的描述**(indices to trace the evolution);
- **社会结构的描述**(indices to test the social structure);

#### ■ **重点**：如何令循环、选择等计算机语言的操作具有特定意义

**Key point:** how to make those computer's operations meaningful for specific research problem;



## 如何抽象语言：规则系统 (rule-based system)

### ■ 语义(semantics):

- “Predicate<Agent, Patient>”: e.g., “吃<狼, 肉>”, “追<狼, 羊>”
- “Predicate<Agent>”: e.g., “跑<狼>”;

### ■ 语言规则（条件+强度）(Linguistic rules):

#### a) 词汇规则(Lexical rules): 语义-语音对(M-U mappings) :

1) 整合词汇规则: e.g., “跑<狼>” $\leftrightarrow$ /a b c/ (0.4)

2) 组合词汇规则:

a) 词: e.g., “吃<#, #>” $\leftrightarrow$ /d e/ (0.3) or “肉” $\leftrightarrow$ /c/ (0.5)

b) 短语: e.g., “吃<狼, #>” $\leftrightarrow$ /c \* f/ (0.4).

#### b) 词序规则(Word order rules): 从普遍能力(排序能力)中发展而来

“Predicate<Agent>”: **SV, VS**

“Predicate<Agent, Patient>”: **SVO, SOV, OSV, VSO, VOS, OVS**

(**S**: Agent 语音音节; **V**: Predicate 语音音节; **O**: Patient 语音音节)

“吃<狼, #>” $\leftrightarrow$ /c \* f/ (0.4)

“肉” $\leftrightarrow$ /c/ (0.5)

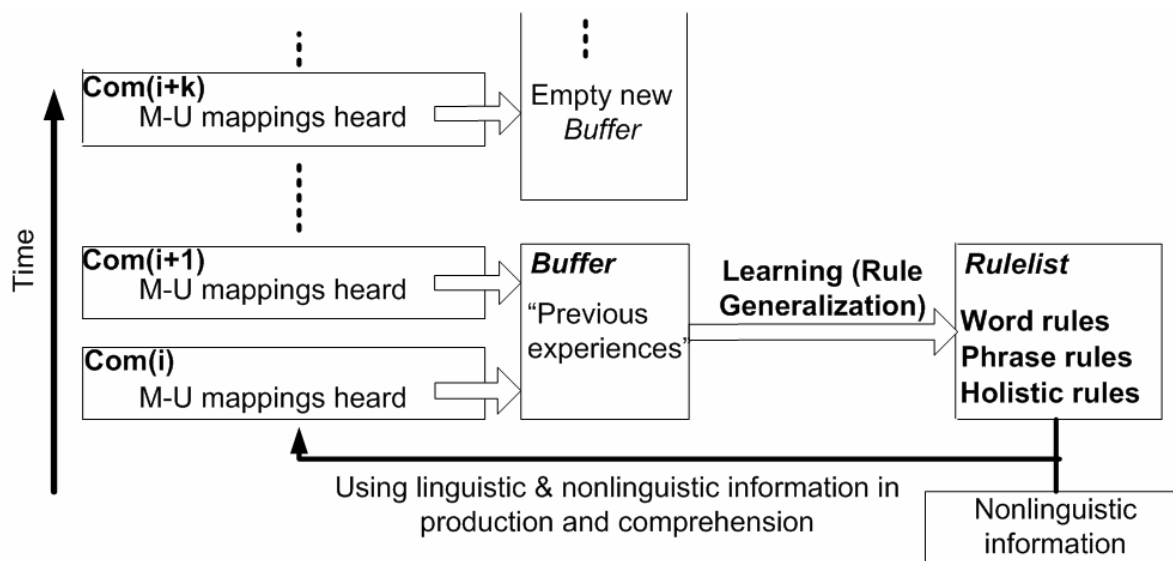


“吃<狼, 肉>” $\leftrightarrow$ /c c f/ (0.4)

**SOV or VOS**

## 个体及其行为:

1) 规则的存储(memory): 两级记忆体 (由分类系统(Classifier System) (Holland 2001)改进而来)



缓存(Buffer): 短期记忆(short-term memory) 以前的经验;

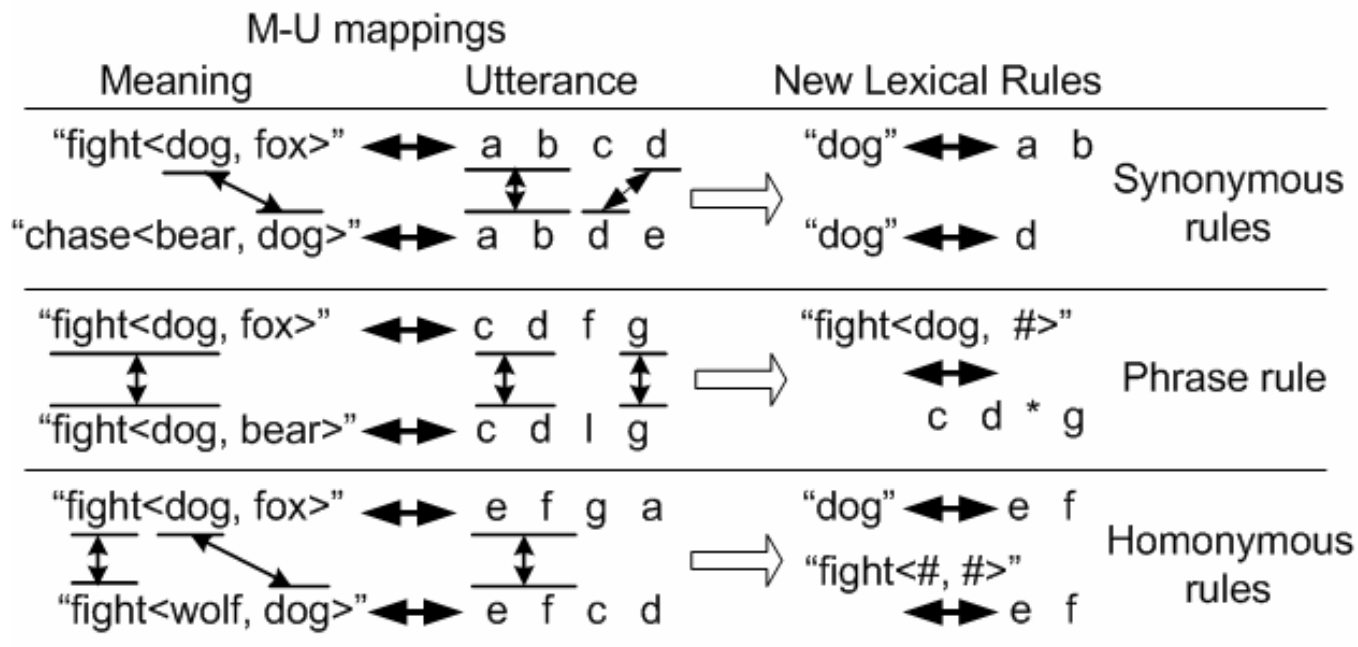
规则表(Rule list): 长期记忆(long-term memory), 存储习得的语言规则并用于交流

## 个体及其行为:

### 2) 规则获得(Rule Acquisition mechanisms)

- **随机创造(Random creation):** 在一定条件下, 对无法用当前规则表示的部分的随机创造;
- **对重复出现的模式的提取(detection of recurrent patterns);**

在个体的缓存中, 当缓存满的时候:

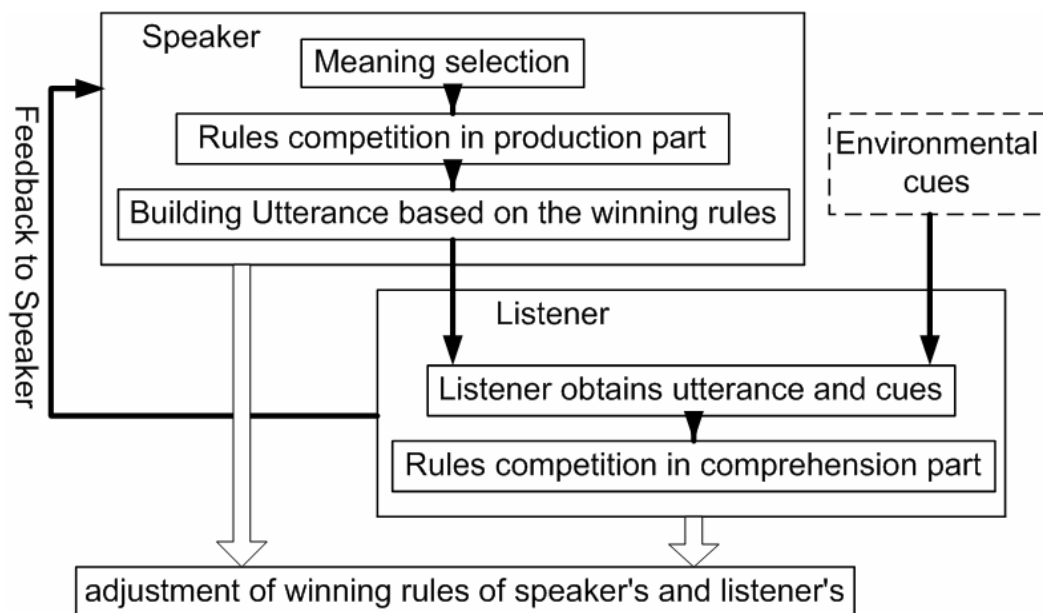


### 3) 交流规则竞争 (Communication and rule competition)

#### ■ 环境信息(Cues):

- 完整语义+强度：理解的提示；
- 所有提示强度相同；
- 环境提示的可信度(Reliability of Cue): 某个提示含有说者语义的可能性；

#### ■ 交流过程图:



# 个体及其行为:

## 3) 交流和规则竞争(Communication and rule competition)

### ■ 程序算法示例:

#### 1) 说者根据现有规则情况确定 (rule selection)

- 确定现有规则可以表达的语义部分;
- 对各种情况的遍历;

#### 2) 规则间基于强度的竞争

(strength-based competition in production and perception)

| Production Part | Meaning to express: "fight<dog, fox>"                                 |                             |  | Utterance created |
|-----------------|---|-----------------------------|--|-------------------|
|                 | Activated rules   | Applicable word order rules | Combined Strength (CS)                 |                   |
| 1 holistic rule | "fight<dog, fox>" ↔ /a b/ (0.6)                                       |                             | CS1 = 0.6                              | /a b/             |
| 3 word rules    | "dog" ↔ /b/ (0.8)<br>"fight<#, #>" ↔ /c e/ (0.5)<br>"fox" ↔ /g/ (0.2) | VSO (0.6)                   | CS2 = 1/2(1/3(0.8+0.5+0.2)+0.6) = 0.55 | /c e b g/         |
| 1 word rule     | "dog" ↔ /b/ (0.8)   | VSO (0.6)                   | CS3 = 1/2(1/2(0.8+0.8)+0.6) = 0.7      | /e b f/           |
| 1 phrase rule   | "fight<#, fox>" ↔ /e * f/ (0.8)                                       | OSV (0.5)                   |  |                   |

| Comprehension Part | Utterance heard: /e b f/  | Cues acquired: "eat<dog, meat>" (0.5) "run<fox>" (0.5) |   | Intended meaning  |
|--------------------|---|--|---|-------------------|
|                    | Activated rules   | Related cues   | Detectable word order rules<br>Combined Strength (CS) |                   |
| 1 holistic rule    | "eat<dog, meat>" ↔ /e b f/ (0.4)                                    | "eat<dog, meat>" (0.5)                                 | CS1 = 1/2(0.5)+1/2(0.5) = 0.45                        | "eat<dog, meat>"  |
| 3 word rules       | "cat" ↔ /e/ (0.7)<br>"fight<#, #>" ↔ /b/ (0.8)<br>"dog" ↔ /f/ (0.6) |  | SVO (0.6) CS2 = 1/2(1/3(0.7+0.8+0.6))+1/2(0.6) = 0.65 | "fight<cat, dog>" |
| 1 word rules       | "run<#>" ↔ /b f/ (0.3)  | "run<fox>" (0.5)                                       | CS3 = 1/2(0.3)+1/2(0.5) = 0.4                         | "run<fox>"        |

Interpreted Meaning: "fight<cat, dog>"



## 演化过程的描述 (indices to trace the evolution)

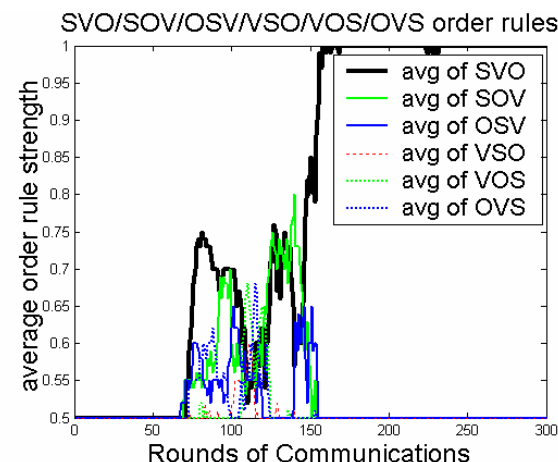
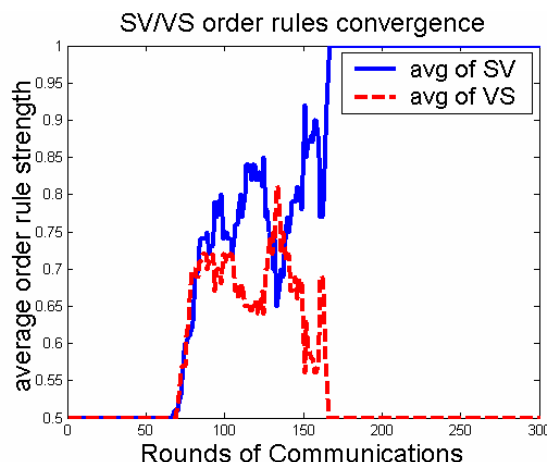
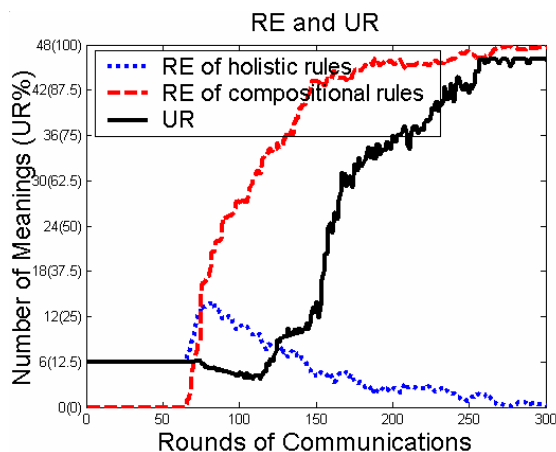
1. 规则的代表能力 (*Rule expressivity (RE)*) — 平均每个个体的各种规则所能表达的完整语义的数目:

$$RE = \frac{\sum_i \text{number of meanings that agent } i \text{ can express}}{\text{number of agents}}$$

2. 正确理解率 (*Understanding rate (UR)*) — 平均每对个体间能够利用语言规则正确理解的完整语义的数目:

$$UR = \frac{\sum_{i,j} \text{number of understandable meanings between agent } i \text{ and } j}{\text{number of all possible pairs of } i, j}$$

## 演化过程的描述 (indices to trace the evolution)



Simulation condition: 10 agents, 500\*5 communications, RC=0.8

### ■ 词汇-语法共同演化 (Lexicon-syntax coevolution)

**驱动力(Driving force):** 相互理解(Mutual understanding)和环境提示(environmental cues);

**共同演化(Coevolution):** 组合词汇规则的使用引发主导词序的收敛, 后者同时促进了组合词汇在个体间的扩散;

### ■ 结论 (conclusion)

- 从整合信号系统向合成语言转变的过程是词汇和语法的共同演化过程。
- 简单语法(如主导词序)可由普遍能力中演化而来。



## 基于交流的简单社会结构的产生 (emergence of simple social structure based on communication)

### 基本假设:

个体为节点(node), 个体间联系为节点间的连接(edge); 通过个体间交流改变连接的权(weight); 当权到达一定值时, 永久连接建立。

社会结构的描述 (indices to test the social structure) (Newman 2003; 刘涛等, 2005)

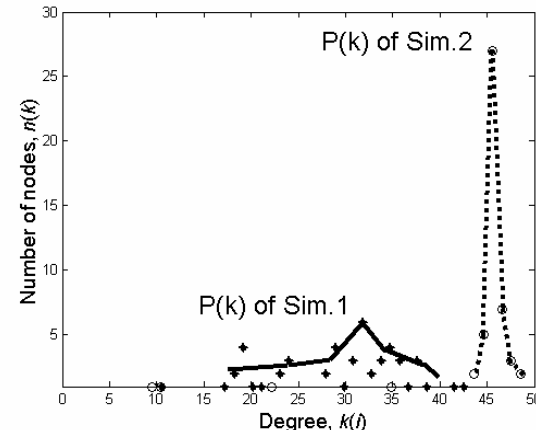
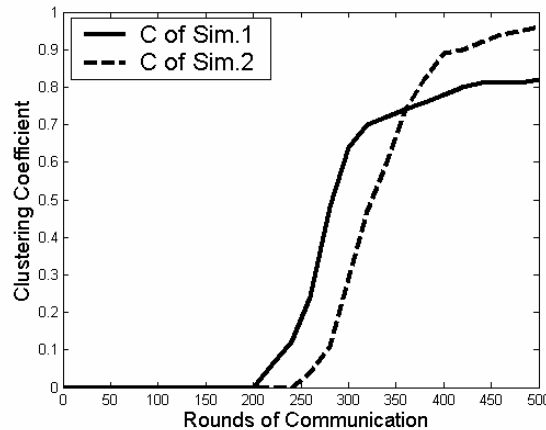
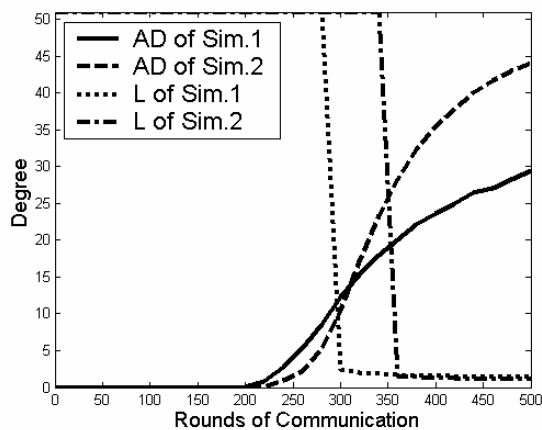
1. 度分布 (Degree distribution ( $P_k$ )): 连接为K的节点占全部节点的分布;
2. 聚集系数 (Clustering coefficient ( $C$ )): 用来描述网络中节点间的聚集情况 (和你相连的节点间有多大可能也存在连接);
3. 平均路径长度 (Average shortest path length ( $L$ )): 所有节点对之间最短距离的平均值, 描述网络中节点间的分离程度;

# 基于交流的简单社会结构的产生

(emergence of simple social structure based on communication)

## ■ 社会结构的产生 (emergent social network):

- $C$  值很高,  $L$  值很小, 说明产生的社会网络具有小世界特性 (Small-World Network) (Watts 1998);
- $P_k$  表明此网络的度分布不具备幂率分布, 说明产生的网络不是无标度网络 (Scale-free Network) (Barabási and Albert 1999);



Simulation condition: 50 agents, 500\*25 communications, RC=0.8, AdjW=0.01, ThresW=0.5

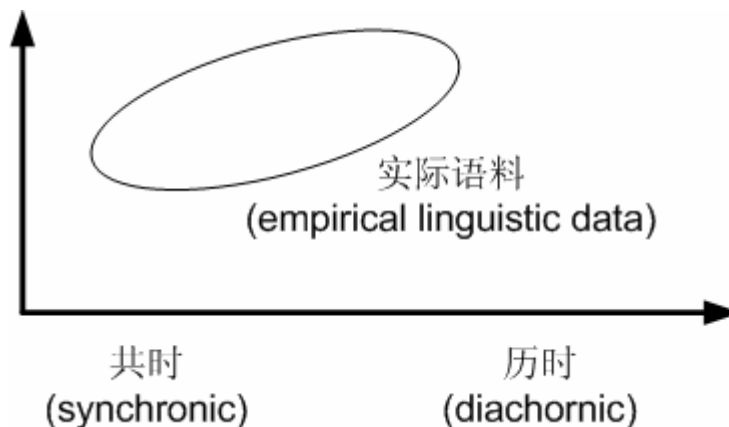
## ■ 结论 (conclusion):

伴随语言的产生, 简单的具有小世界特性的社会结构也同时产生.

## 讨论与总结

### Discussion and Conclusion

- 模型设计注意事项：  
(crucial points in model design)
  - 假设、参数及其合理性  
(assumption, parameters and their validities)
  - 结果采集与相关分析 (analysis of results)
  - 实际数据的匹配 (matching empirical data)



## 讨论与总结

### Discussion and Conclusion

- 相关演化方面的语言学课题 (other evolutionary linguistic topics):
  - 语言起源(language emergence);
    - 群体语言历时演化(phylogenetic emergence of language);
    - 个体语言共时习得(ontogenetic emergence of language);
  - 语言发展(language change);
    - 语言接触(language contact);
    - 社会结构对语言创新传播的影响(linguistic innovation and social structure);
  - 语言转用与消亡(language death);
    - 由接触导致的语言转用：倒话，瓦乡话；
    - 濒危语言(endangered languages)：少数民族语言：白语，畲语。



## 讨论与总结

### Discussion and Conclusion

- **模拟仿真模型 (computational simulations):**
  - 根据模型的目的划分(separation by objectives (Holland 2005))
    - **数据驱动模型(data-driven):** 气象预报模型(weather prediction model);
    - **证明可存在模型(existence-proof):** 自复制机模型(self-reproduction machine);
    - **探索性模型(exploratory):** 我们的模型(our model);
  - 根据所模拟和所使用的工具划分(separate by action and tools adopted)
    - **行为模型(behavior-based modeling):** e.g., 多个体模型、神经网络模型;
    - **数学分析模型(mathematical-based modeling):** e.g., 马可夫链、动态分析模型、语言关系树;
  
- **其他应用领域 (other applications of computational simulation):**
  - 优化问题(optimization);
  - 人工智能(artificial life/intelligence)
  - 其他演化相关问题(other evolutionary phenomena)

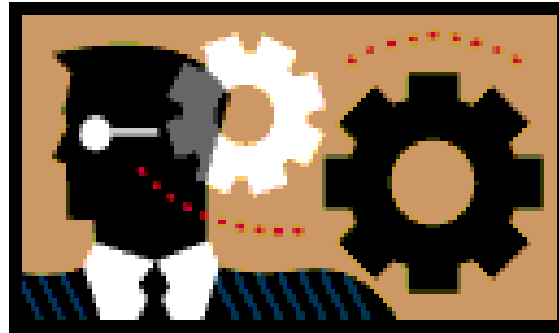
## 参考文献 (Selected references)

- Barabási, A. L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286: 509–512.
- Calvin, W. H. and Bickerton, D. (2000). *Lingua ex machine: reconciling Darwin and Chomsky with the human brain*. Cambridge, MA: MIT Press.
- Chomsky, N. (1995). *The Minimalism Program*. Cambridge, MA: MIT Press.
- Christiansen, M. and Kirby, S. eds. (2003). *Language Evolution*. Oxford, NY: Oxford University Press.
- Hasuer, M. D. (1996). *The Evolution of Communication*. Cambridge, MA: MIT Press.
- Holland, J. H. (2001). Exploring the Evolution of Complexity in Signaling Networks. *Complexity*, 7(2): 34–45.
- Holland, J. H. (2005). Language Acquisition as a Complex Adaptive System. *Language Acquisition, Change and Emergence – Essays in Evolutionary Linguistics*. Hong Kong: City University of Hong Kong Press.
- Jackendoff, R. (1997). *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.
- Kirby, S. (2002). Learning, Bottlenecks and the Evolution of Recursive Syntax. *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge, MA: Cambridge University Press.
- Knight, C., Hurford, J. R. and Studdert-Kennedy, M. eds. (2000). *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*. Cambridge, MA: Cambridge University Press.
- Munroe, S. and Cangelosi, A. eds. (2002). *Simulating the Evolution of Language*. London, New York: Springer.
- Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45(2): 167–256.

- Smith, K., Brighton, H., and Kirby, S. (2003). Complex Systems in Language Evolution: the Cultural Emergence of Compositional Structure, *Advances in Complex Systems*, 6(4):537–558.
- Wagner, K., Reggia, J. A., Uriagereka, J., and Wilkinson, G. S. (2003). Progress in the simulation of emergent communication and language. *Adaptive Behavior*, 11(1):37–69.
- White, T. D. et al. (2003). Pleistocene *Homo sapiens* from Middle Awash, Ethiopia, *Nature* 423(12).
- Wang, W. S-Y. (1978). *Diamond Jubilee Lecture*. Osmania University. Dec.1978.
- Wang W. S-Y. (1991). The Three Scales of Diachrony. *Explorations in Language*, ed. by Wang, W. S-Y, 60–71. Taiwan: Pyramid Press.
- Wray, A. (2002). *Formulaic Language and the Lexicon*. New York: Cambridge University Press.
- 刘涛, 陈忠, 余哲. (2005). 复杂网络理论及其应用研究概述. 中国科技论文在线 [www.paper.edu.cn](http://www.paper.edu.cn).

## 本小组相关论文 (Selected papers of our group)

- Ke, J., Minett, J. W., Au, C. P. and Wang, W. S-Y. (2002). Self-organization and selection in the emergence of vocabulary. *Complexity*, 7(3):41–54.
- Gong, T., Ke, J., Minett, J. W. and Wang, W. S-Y. (2004). A Computational Framework to Simulate the Coevolution of Language and Social Structure. In: Pollack, J., Bedau, M., Husbands, P., Ikegami, T. and Watson, R. A., eds., *Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems (ALIFE9)*, 158–163, Cambridge, MA: MIT Press.
- Gong, T., Ke, J., Minett, J. W., Holland, J. H. and Wang, W. S-Y. A Computational Model of Coevolution of Lexicon and Syntax. *Complexity*, Submitted.
- Gong, T. and Wang, W. S-Y. (2005). Computational Modeling of Language Emergence: Coevolution of Lexicon, Syntax and Social Structure. *Language and Linguistics*, 6(1): 1–41.



*Thank You*