Visit to The University of Oxford:

Linguistic problems from simulation perspective

– A lexicon-syntax coevolution model

GONG Tao
tgong@ee.cuhk.edu.hk
LEL and DSP Lab.,
Dept. of EE, The Chinese University of Hong Kong
Aug.16, 2005;

Outline

- Background;
  - Research methodology – computational simulation;
  - Hypotheses on language emergence;
- Overview of the lexicon-syntax coevolution model;
- Simulation results;
  - Indices to test the performance;
  - Lexicon-syntax coevolution;
- Something about our group;

Background

Research methodology on linguistic problems

- *Traditional approaches:*
  - Anthropology & Archaeology findings *(e.g., White et al. 2003)*;
  - Cognitive & Neuro-science *(e.g., Calvin and Bickerton 2000)*;
  - Empirical linguistic data collection (e.g., CHILDES, Switchboard, etc.);

- *Newly developed approaches:*
  - Computational Simulation:
    - Hawkins, J.A. & Gell-Mann, M., eds. 1992. *The Evolution of Human Languages*. Addison-Wesley.
    - Hurford, J.R., Studdert-Kennedy, M., Knight, C., eds. *Approaches to the Evolution of Language.* Cambridge University Press, 1998.
    - Cangelosi, Angelo & Parisi, D., eds. 2002. *Simulating the Evolution of Language.* Springer.
    - Christiansen, M.H. & Kirby, S., eds. 2003. *Language Evolution.* Oxford University Press.
    - Minett, J.W. & Wang, W. S-Y., eds. 2005. *Language Acquisition, Change and Emergence.* Hong Kong: City University Press.

Model = Theory = Framework

- Types of model:

  Based on objectives (*Holland,* 2005)*:*

  - **Predictive model** — generates output that mimics and predicts empirical data: **weather forecast; comparative method** …
  - **Proof of principle model** — proves that something is possible **in principle**: (von Neumann's self-replicating machine) …
  - **Exploratory model** — explore the possibilities in a set of mechanisms: (Conway's Game of Life) …

  Based on approaches:

  - **Top-down model** — typically a predictive model that operates on empirical data to **infer the processes** of interaction that process the data;
  - **Bottom-up model** — a model that operates on basic elements and the interactions among them to **infer their collective behavior;**

  Based on methods:

  - **Deterministic models** — Mathematics-based methods: Markov Chain, Differential Equations; **Language death, lexical diffusion**;
  - **Probabilistic, behavior-based methods** — e.g., **Multi-agent model**, Neural networks;

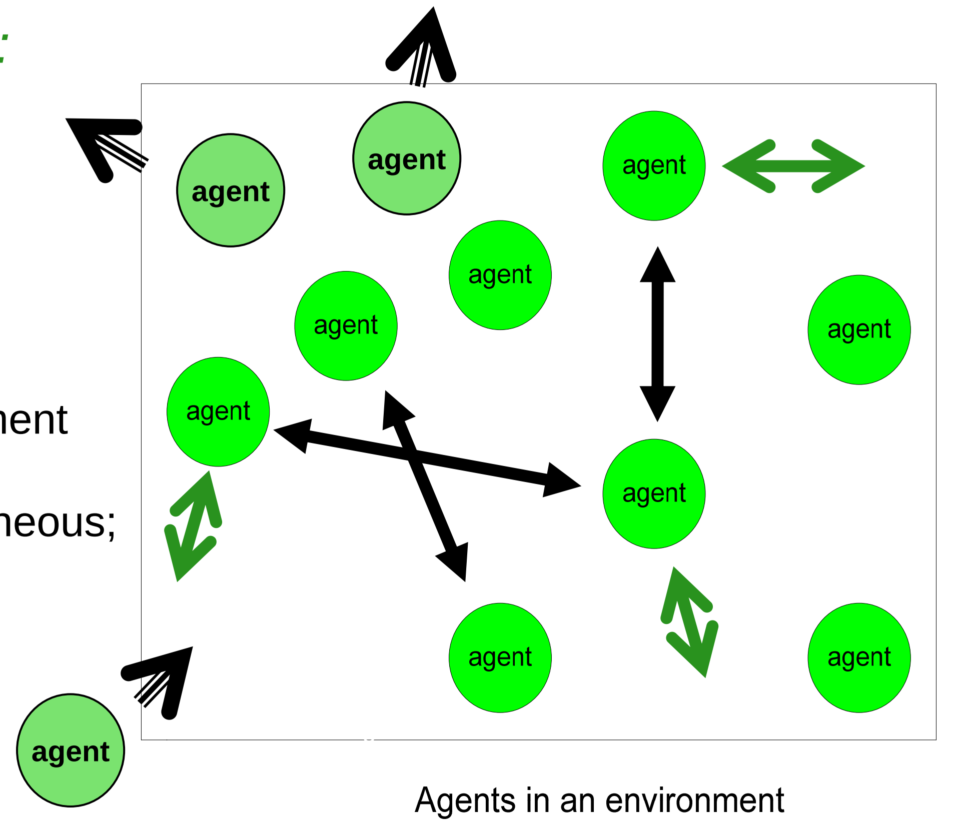
Multi-agent system (according to *Steels* 2004; *Wagner* et al. 2003)

Components of multi-agent system:

Agent:
1) Independent, automatous unit;
2) Possess certain abilities;

Agent's abilities:
1) Memory: e.g., rule-based system;
2) Activities: e.g., communication, movement and replacement;
3) Agents are homogeneous or heterogeneous;

Environment:
1) Situated (actions between agents and environment);
2) Non-situated;



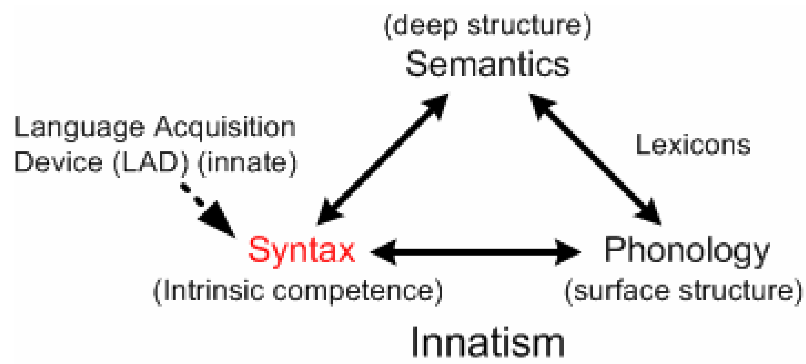
Agents in an environment

Aims of multi-agent system:

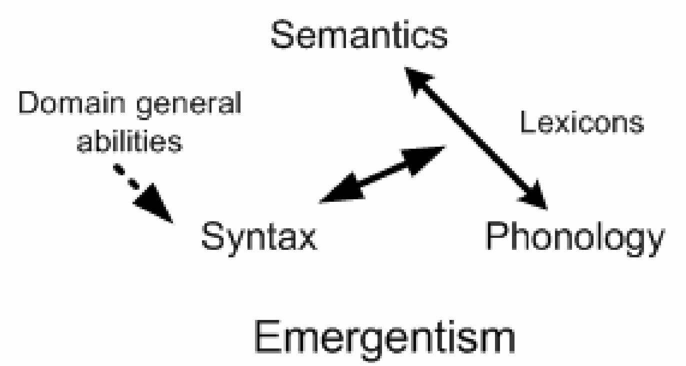
After certain rounds of activities, test the emergence of certain global tendencies:
1) Share common rules;
2) Possess common behaviors;
3) Relationships among agents, and between agents and environment;

Hypotheses on language emergence

- *Innatism* (*Chomskyan school*)
  - Children do not "learn" a language, as such, but use what they hear to set the **parameters** of a **Universal Grammar** (UG) in order to understand and speak a particular language.

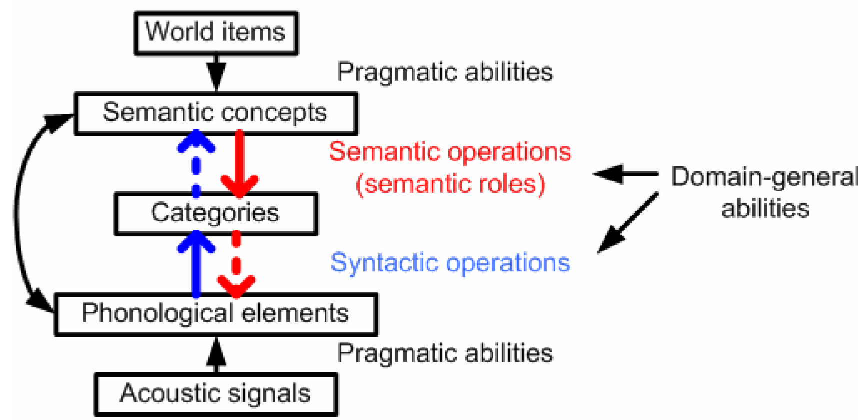


- *Emergentism* (e.g., *Hurford et al.* 2000; *Oller and Griebel* 2004)
  - "[Language] evolved in a mosaic fashion, with the emergence of semantics, phonology, morphology and syntax all at different times and according to different schedules ... language is regarded as a kind of 'interface' among a variety of more basic abilities. These abilities underlie nonlinguistic processes as well." (*Wang 1978*)



Lexicon-syntax coevolution model

- *Target Questions:*
  - Could linguistic competences be derived from some domain-general abilities?
  - Instead of built-in, could syntax be gradually developed during the acquisition of lexicons?
- *Research methodology:*
  - Multi-agent computational model;
  - Rule-based system to represent linguistic information;
  - Evolutionary computation mechanisms;
- General assumptions:
  - No global word orders built-in, they are gradually formed based on local orders between two categories;
  - Linguistic knowledge is gradually formed based on semantic and syntactic operations, both of which are derived from domain-general abilities;



Components of the model:
Meaning-Utterance (Meaning-Form) mappings

- *Integrated meanings:*
  - “Predicate<Agent>”: e.g., “run<wolf>”;
  - “Predicate<Agent, Patient>”: e.g., “chase<wolf, fox>”, “eat<lion, meat>”;

- *Utterance:* Combinable syllables from a syllable space;

- *Meaning-Utterance mappings:*
  - Holistic mapping: the utterance as a whole is mapped to some integrated meaning;
    - e.g., “run<wolf>”←→ /e f g/
      “fight<bear, wolf>” ←→ /c g/
  - Compositional mapping: the meaning mapped by the utterance is a function of the utterance components;
    - e.g., “run<wolf>” ←→ /a g c/
      “chase<wolf, bear>” ←→ /a b c d/

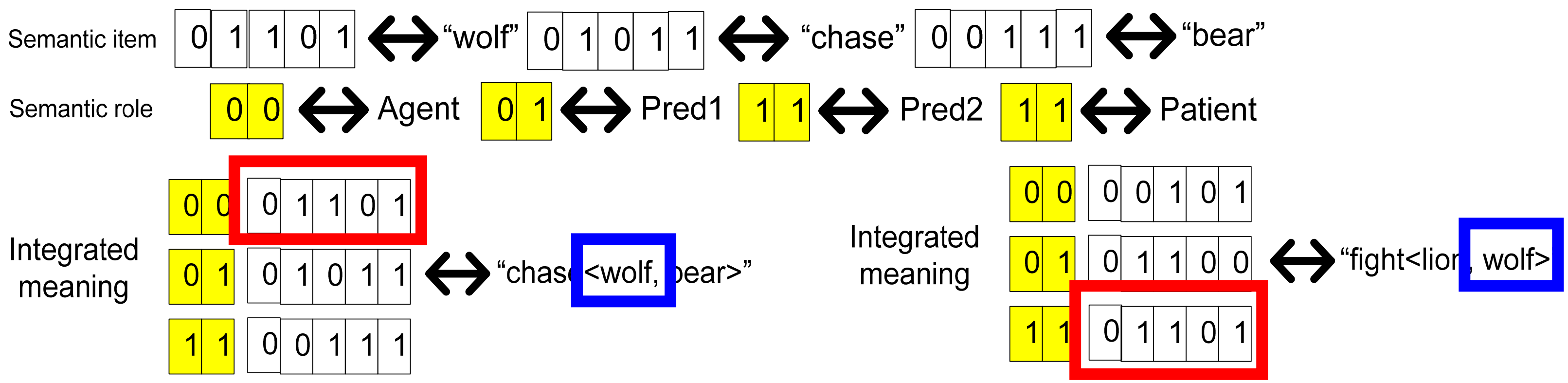
Semantics operation **Syntactic operation**

↘ ↙

Category (linguistic knowledge)

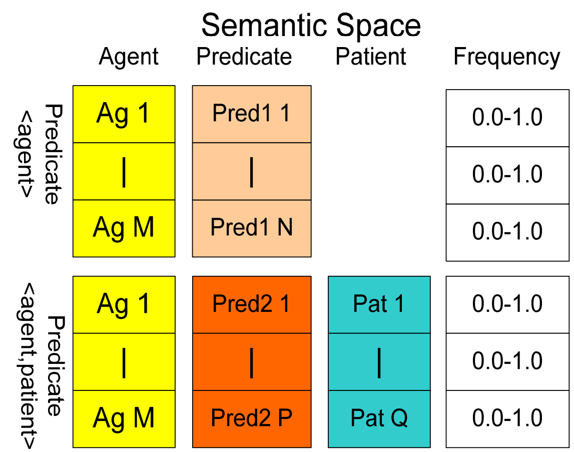
Semantic operation

- Semantic items, semantic roles and integrated meanings;



Semantic space: integrated meanings + frequency (*personal experience*);

Semspace1: M=4, N=4, P=4, Q=4 (="Ag"):
"Pred1<Ag>": 4*4=16;
"Pred2<Ag, Pat>": 4*4*3=48;



Syntactic operation and syntax

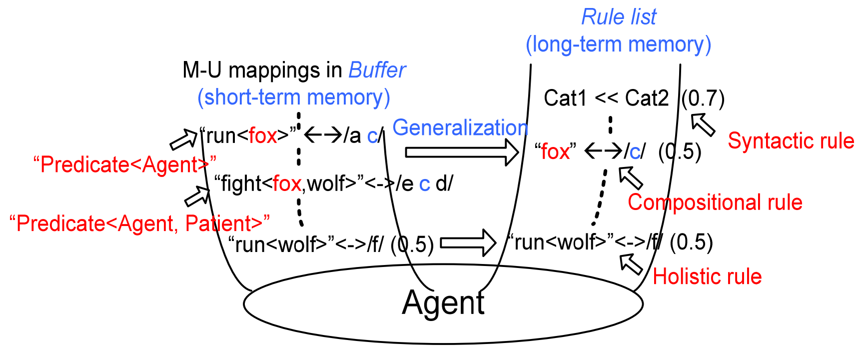
- Syntactic roles of compositional utterance syllables:
  - Semantic category $\neq$ syntactic category;
  - Subject (S) (“Agent”), Verb (V) (“Pred1”,”Pred2”), Object (O) (“Patient”);

- Local order: order relations between 2 utterance items with different syntactic roles.
  - Orders: << (before) and >> (after); ▲ (middle) and ▼ (surround)
  - /a b/ << /c/ ←→ Verb << Subject (VS);

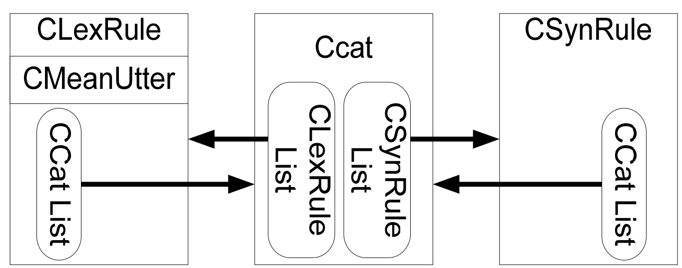
- Syntactic operation:
  - Generalize local orders among different utterances;
    - “run<wolf>” ←→ /a g c/; “chase<wolf, bear>” ←→ /a b c d/
  - Build up global orders based on local orders;
    - unique order: V << S + O >> S ➔ VSO;
      multiple orders: V << S + O << S ➔ VOS or OVS;
      SV ▼ O ➔ SOV or VOS;

Linguistic rules and category

- Linguistic rules: (condition + strength);



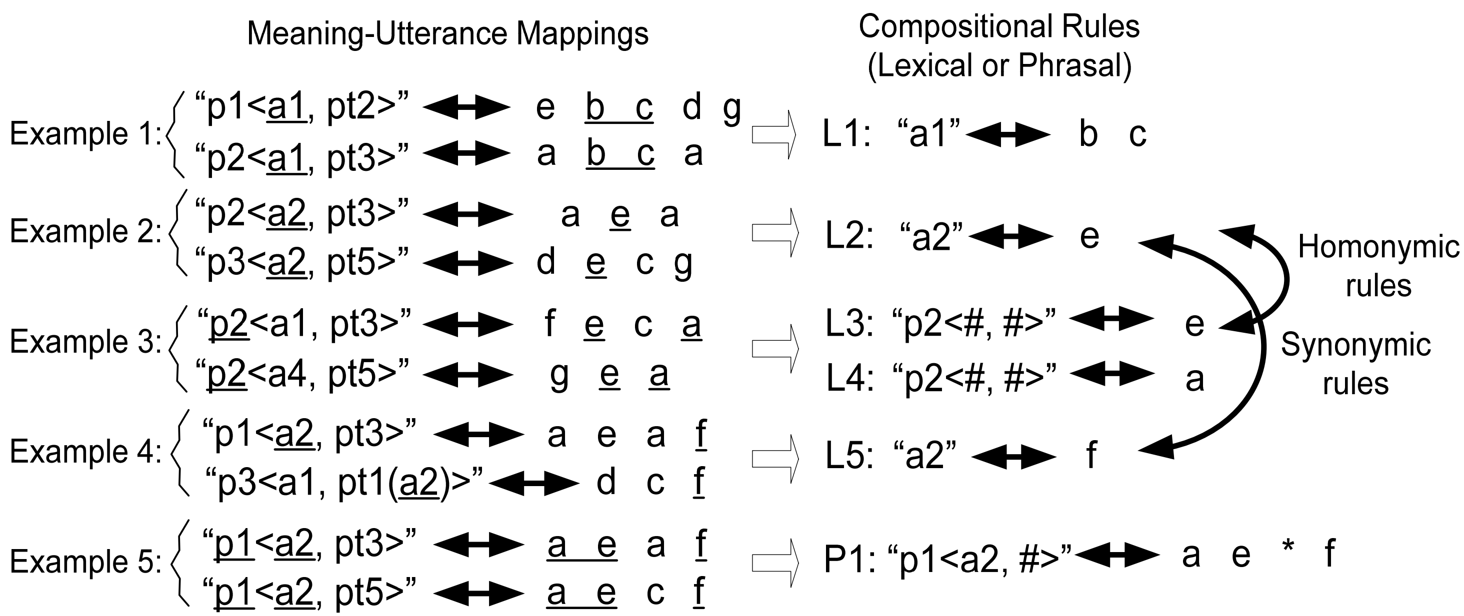
- Category (lexical rule list + syntactic rule list):
  - Knowledge of how to use lexical rules and syntactic rules to produce/comprehend integrated meaning;



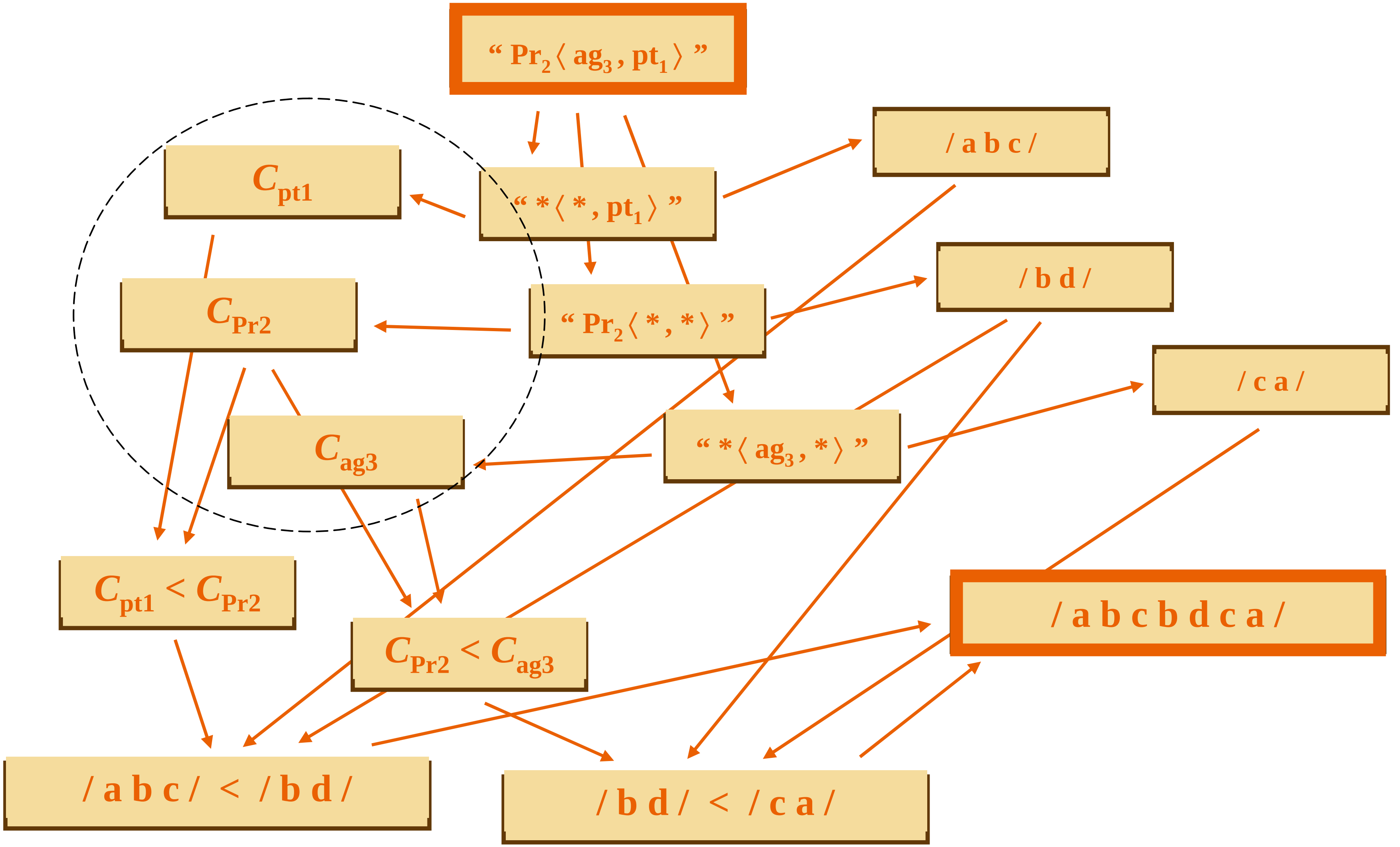
  - Operations on category:
    1) Category formation: lexical rules following same local orders can form a category;
    2) Category merge: if 2 categories 1) share same syntactic roles and some lexical rule members, and 2) toward lexical rule members of another category, the shared members follow same local order, they can merge into one category;

Components of the model:
Rule acquisition mechanisms;

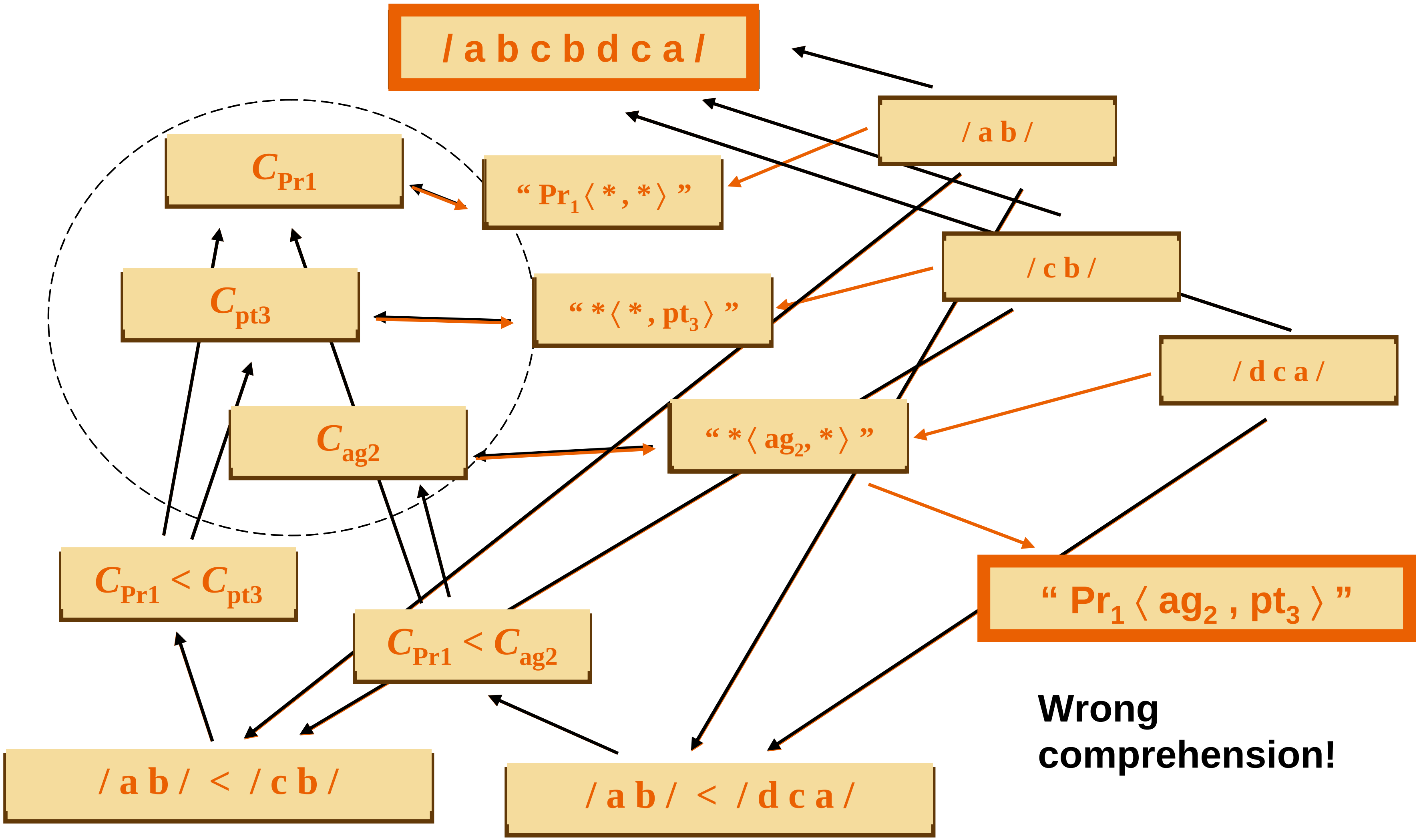
- Rule acquisition mechanisms:
  - Random creation (similar to Iterative Learning Model (*Kirby* 2002);
  - Detection of recurrent patterns: simplified pattern-recognition;
  - Synonyms and homonyms are possible;



Communication: Speaker's production



Communication: Listener's perception



Wrong comprehension!

Simulation Results

- Simulation scenario:
  - From holistic signaling system to compositional language (*Wray* 2002);
  - **Initial stage:** holistic signaling system, all agents share some common holistic rules, with no syntactic rules nor categories;
  - **Random communication:** On each step, 2 randomly chosen agents communicati with each other;

- Indices to test the performance:
  - Rule expressivity (RE): "how many meanings on average one agent can express":

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

  - Understanding rate (UR): "how many meanings on average one agent can accurately understand based on linguistic information only (displacement (*Hockett* 1960))":

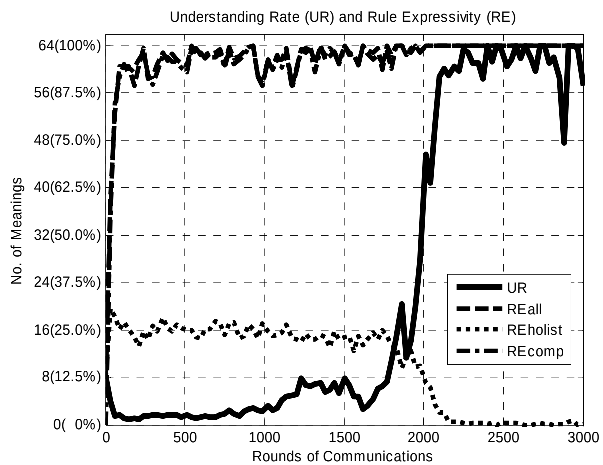
$$UR = \frac{\sum_{i,j} \text{number of understable meanings between agent i and j}}{\text{number of all possible pairs of i, j}}$$

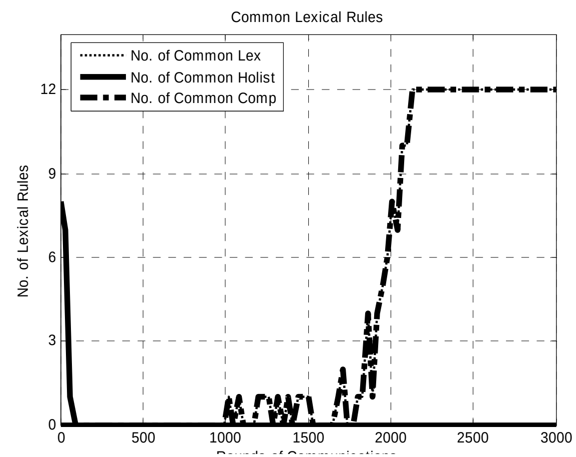
  - Number and type of common lexical rule;
  - Global word orders' productivity and understandability;

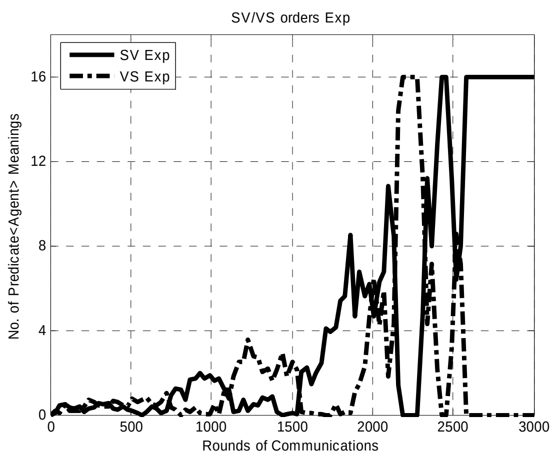
Simulation Results: Lexicon-Syntax Coevolution

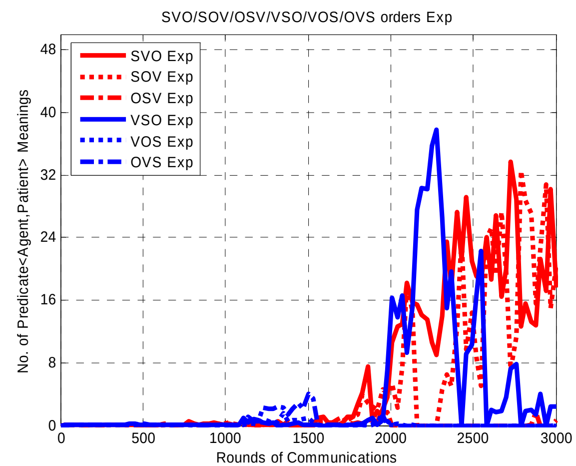
- Simulation conditions:
  - N_{Agent}=10; N_{Com}=3,000; $N_{MeanExc}$=20; S_{buffer}=30; S_{Lex}=40; S_{Cat}=20; S_{Syn}=20;
  - Strength (Lex-rule/Syn-rule str, Asso-Weight): [0.0 1.0]; InitStr=0.5; AdjStr=0.1; ForgetStr=0.01;
  - CueReliab=0.8; Confidence Threshold=0.75;
  - Semantic Space: 64 (16+48) Mean; Type ratio = 1:3; **Token ratio = 3:1;**

Lexicon-Syntax coevolution

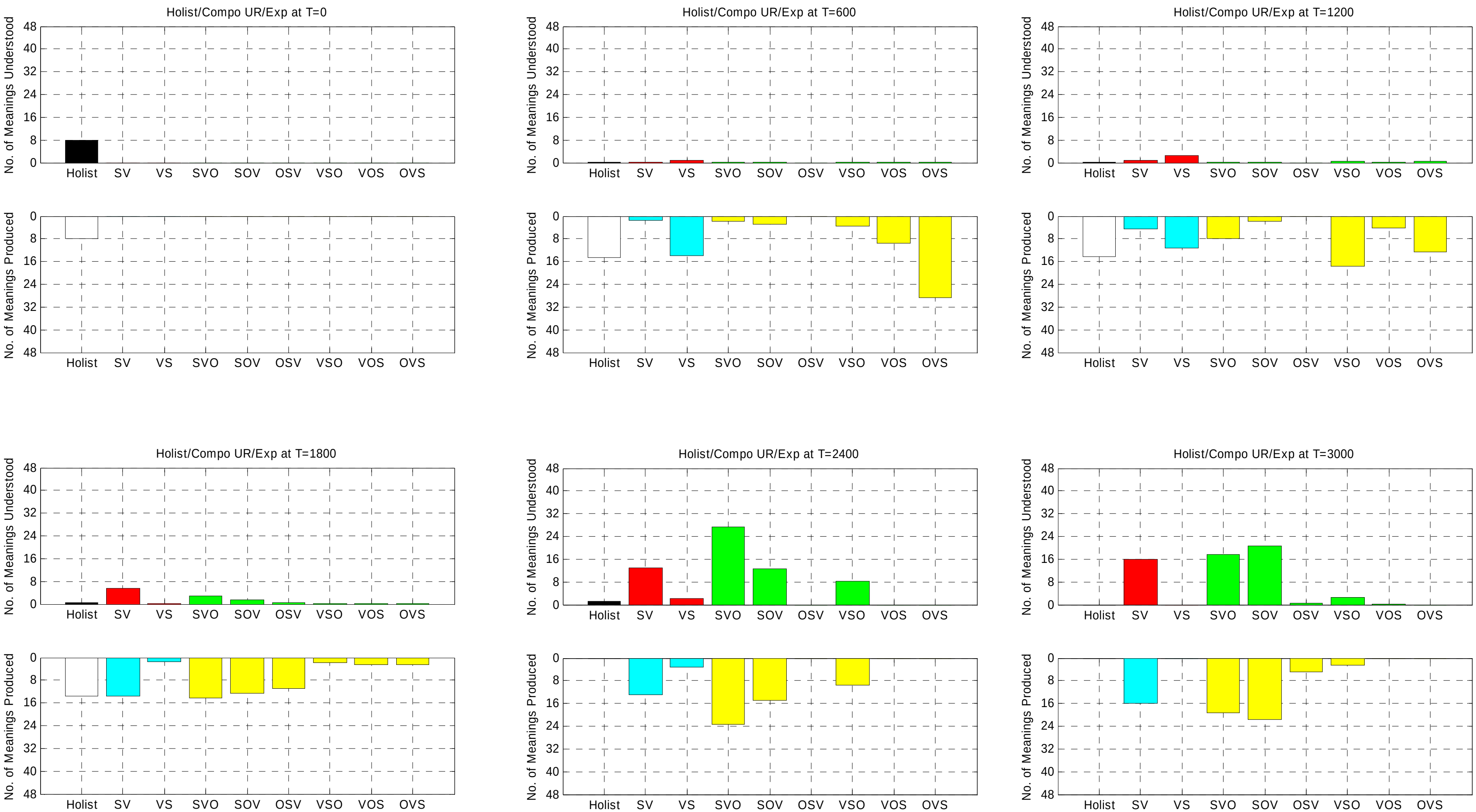








Simulation Results: Coexistence of multiple global orders



Explanation of coexistence of multiple orders

- Semantic category ≠ syntactic category;
  - Modern languages with flexible orders, e.g., French.
- Instance frequency's influence on creation of local orders:
  - Inst ("Ag") > inst ("Pred1") = inst ("Pred2") > inst ("Pat") ➔ SV/VS >> SO/OS >> VO/OV;
- Parsimony in using local orders in production and perception:
  - **Multiple orders are possible:** In production, 2 relative orders are enough to produce stable/multiple global orders;
    - e.g., V << S + O >> S ➔ VSO; V << S + O << S ➔ VOS or OVS;
  - **Multiple orders are stable:** In perception, 2 relative orders are enough to distinguish S, V and O and get an accurate comprehension;
    - e.g., utterance with VOS or OVS global order can be accurately comprehended using local orders: V << S + O << S;
  - Examples of language with flexible orders: e.g., French, Chinese;

Conclusions

- Preliminary results:
  - The transition from holistic signaling system to compositional language is a conventionalization process of acquiring linguistic knowledge (categorization).
  - This transition is a coevolutionary process of the emergence of common lexicons and similar global orders (both of which are achieved within same period);
  - Understanding rate follows a "S-curve" with a sharing increasing stage;
- Possible cooperation fields:
  - Empirical neuro-experiments ➔ computational simulation
    - Empirical basis for simulation assumptions;
    - Mechanisms for simulation;
    - Test simulation results;
  - Computational simulation ➔ empirical neuro-experiments
    - Test mechanisms summarized from empirical experiment;
    - Suggest/predict empirical experiments;

Something about LEL

http://dsp.ee.cuhk.edu.hk/lel/;
first founded in 1994 in City U of HK; moved to CUHK in Sep. 2004.

- **Supervisor:** Prof. Wang, W. S-Y.;
- **Pos-doc:** Dr. Minett, J. W., Peng, G.;
- **Current members:** Wong, F., Gong, T., Shuai, L., Zheng, H. Y.;
- **Evolutionary linguistic topics in LEL:**
  - Language emergence;
    - Phylogenetic emergence of language (*Gong, T & Minett, J.*);
    - Ontogenetic emergence of language (Wong, F. & Au, B.);
  - Language change;
    - Language contact (*Ke, J.*); Linguistic innovation & social structure (*Gong, T., Minett, J. & Ke, J.*);
  - Language death;
    - Language competition on users; Endangered languages (*Minett, J. & Wang, F.*);

Minett, J. W. & Wang, W. S-Y., eds. (2005) *Language Acquisition, Change and Emergence: Essays in Evolutionary Linguistics.* Hong Kong: City University of Hong Kong Press.

Gong, T., Ke, J., Minett, J. W., Holland, J. H. & Wang, W. S-Y. (2005) “A computational model of coevolution of lexicon and syntax”, Complexity 10.6: 50-62.

Gong, T. & Wang, W. S-Y. (2005) “Computational modeling of language emergence: coevolution of lexicon, syntax and social structure”, Language and Linguistics 6.1: 1-42.

Gong, T., Minett, J. W. & Wang, W. S-Y. (2005) “Computational exploration on language emergence and cultural dissemination”, IEEE Congress on Evolutionary Computation (IEEE CEC 2005).

Gong, T., Ke, J-Y., Minett, J. W. & Wang, W. S-Y. (2004) “A computational framework to simulate the coevolution of language and social structure”, Ninth International Conference on the Simulation and Synthesis of Living Systems (ALIFE9), Boston, US:158–163.

Ke, J-Y., Minett, J. W., Au, C-P. & Wang, W. S-Y. (2002) “Self-organization and selection in the emergence of vocabulary”, Complexity 7.3: 41–45.

Selected references

Calvin, W. H. & 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. & Kirby, S. eds. (2003) *Language Evolution*. Oxford, NY: Oxford University Press.

Hasuer, M. D. (1996) *The Evolution of Communication*. Cambridge, MA: MIT Press.

Jackendoff, R. (1997) *The Architecture of the Language Faculty.* Cambridge, MA: MIT Press.

Hockett, C. F. (1960) “The origin of speech”, Scientific American 203: 88-96.

Holland, J. H. (2002) “Exploring the evolution of complexity in signaling networks”, Complexity 7.2:34-45.

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. & Studdert-Kennedy, M. eds. (2000) *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*. Cambridge, MA: Cambridge University Press.

Smith, K., Brighton, H. & Kirby, S. (2003) “Complex systems in language evolution: the cultural emergence of compositional structure”, Advances in Complex Systems 6.4:537–558.

Steels, L. (2004) “The evolution of communication systems by adaptive agents”. In: Alonso, E., D. Kudenko and D. Kazakov (eds.) *Adaptive Agents and Multi-Agent Systems*. LNAI 2636. Springer Verlag, Berlin: 125-140

Wray, A. (2002) *Formulaic Language and the Lexicon*. New York: Cambridge University Press.

Wang, W. S-Y. (1978) *Diamond Jubilee Lecture*. Osmania University. Dec.1978.

Wang W. S-Y. (1991) “The three scales of diachrony”, In: Wang, W. S-Y. (ed.) *Explorations in Language*. Taiwan: Pyramid Press: 60-71.

Wagner, K., Reggia, J. A., Uriagereka, J. & Wilkinson, G. S. (2003) “Progress in the simulation of emergent communication and language”, Adaptive Behavior 11.1:37–69.