



Language Contact from a Simulation Perspective

GONG Tao

LEL & DSP Lab.

Dept. of EE, The Chinese University of Hong Kong

May 14, 2007



Outline

- Language contact, and the effects of social context and linguistic features in this process
- The simulation perspective from computational linguistics to explore linguistic problems
- Two simulation studies on intra- and inter-group language contact
- Conclusions and final remarks



Language Contact:

- **Language contact:** the prolonged association between the speakers of different (or the same) languages (Thomason and Kaufman 1988; Crystal 1992);
- Two common situations in language contact:
 - **Borrowing:** the incorporation of foreign features into a group's native language by speakers of that language, e.g., lexical loans from English to Cantonese;
 - **Substratum:** results from imperfect group learning during a process of language shift, e.g., pidgin and creole languages of African slaves in America (Mufwene 2001);
- **The major factors relevant to these situations in language contact:**
 - **Social context:** “the history of a language is not an independent phenomenon that can be thoroughly studied without reference to the social context in which it is embedded” (Thomason and Kaufman 1988).
 - **Linguistic constraints:** lexical items (Swadesh 1952) and syntactic features may affect the process of language contact.

Thomason, S. G. and Kaufman, T. 1988. *Language contact, creolization, and genetic linguistics.*

Crystal, D. 1992. *The Cambridge encyclopedia of language.*

Mufwene, S. S. 2001. *The ecology of language evolution.*

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts, *Proceedings of the American philosophical society* 96(4): 452-463.



Report of an empirical study on Daohua 倒话 (Atshogs 2003)

- A creole language in Yalong river, Southwest Sichuan Province, China, between the Tibetan and Han areas.
- It emerged about 300 years ago during the invasion of the Qing troops into Tibet to suppress the local minority riots.
- It is hybrid of *Southwest Mandarin* and *Tibetan*.
- Now it is documented and studied by *Endangered Languages Documentation and Preservation (ELDP)*.



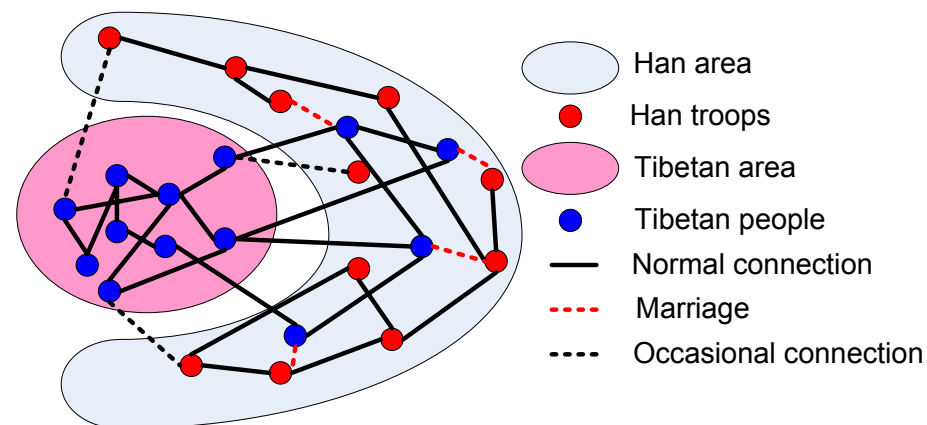
- Some example sentences in *Daohua*:
 - 马 骑 人 一个 来了。 /
 - Horse ride man one come
 - "一个骑马的人来了" (A man riding a horse came).
 - /他 -ki 茶 喝 -ts'e-tsu⁴-di-jiu³-li。 /
 - he (subject) tea drink (when is going to and not begin yet)
 - "他正要喝茶 (还没喝) 的时候" (when he is going to and not yet begin to drink the tea)



Report of an empirical study on Daohua 倒话 (Atshogs 2003)

- The major factors that cause the emergence of *Daohua*:

Social context: the unique social structure in this area causes different **Degrees of Communicative Pressure (DCP)**, in all communicative activities of a language user, the percentage of the activities in which a foreign language is adopted) for Han (**DCP=3**) and Tibetan (**DCP=4**) people;



Contact pattern of *Daohua*:

L: Language; **T:** Tibetan; **M:** Mandarin; **D:** Daohua

V: Vocabulary; **S:** Semantics; **G:** Grammar; **P:** Phonology

$L(D) = L(T) \leftrightarrow L(M) = \{V(M), G(\text{mainly } T), P(\text{mainly } M), S(T\&M)\}$



The simulation perspective

- Complex contact effects are usually **unpredictable** (Thomason 2000):
 - Many common or natural changes don't occur in a particular language at a particular time:
 - e.g., *Montana Salish* (a Northwest Native American language) in the US exposed intensive contact with English, but it has some, but very few loanwords from English, and no structural interference from English either.
- *Besides the empirical studies of particular languages, is there any other method to systematically study the effects of social context and linguistic features on language contact? – Yes!*
- Contemporary linguistic research methods:
 - **Empirical data collections:** e.g., field work and corpus linguistics;
 - **Experimental studies on human subjects' performance in linguistic tasks:** e.g., categorical perception and neural activities during communications;
 - **Computational simulation:**
 - Adopt computational models to **recapitulate** the evolution of human language, **reconstruct** the language history, and **reconsider** the effects of linguistic and nonlinguistic factors on this developmental process.
 - Computational models on language origin, change (via contact), and death.



The advantage of computational simulation

- **Complementary** to empirical studies or theoretical analyses on language evolution.
 - Break through some limitations of the traditional means;
 - Help to verify the incomplete or unreliable linguistic theories or hypotheses.

- Precisely studying the nature of language as a **Complex Adaptive System** (CAS, **Steels 2000**; **Wang 2006**):
 - Through adjusting parameters and repetition under similar conditions, computational simulation can systematically study the collective and general effects of various factors on language evolution.

- As **reliable** as empirical and experimental studies:
 - Adopt plausible assumptions verified by empirical findings in linguistics and other disciplines.
 - Use objective, realistic mechanisms, and follow well-defined, traceable procedures to obtain convincing and replicable results.

Steels, L. 2000. Language as a Complex Adaptive System, *Proceedings of the 6th International Conference on Parallel Problem Solving from Nature*: 17–26.

Wang, W. S-Y. 2006. 语言是一个复杂适应系统 (Language is a CAS). *Journal of Tsinghua University (Philosophy and Social Sciences)*, 21(6): 5-13.

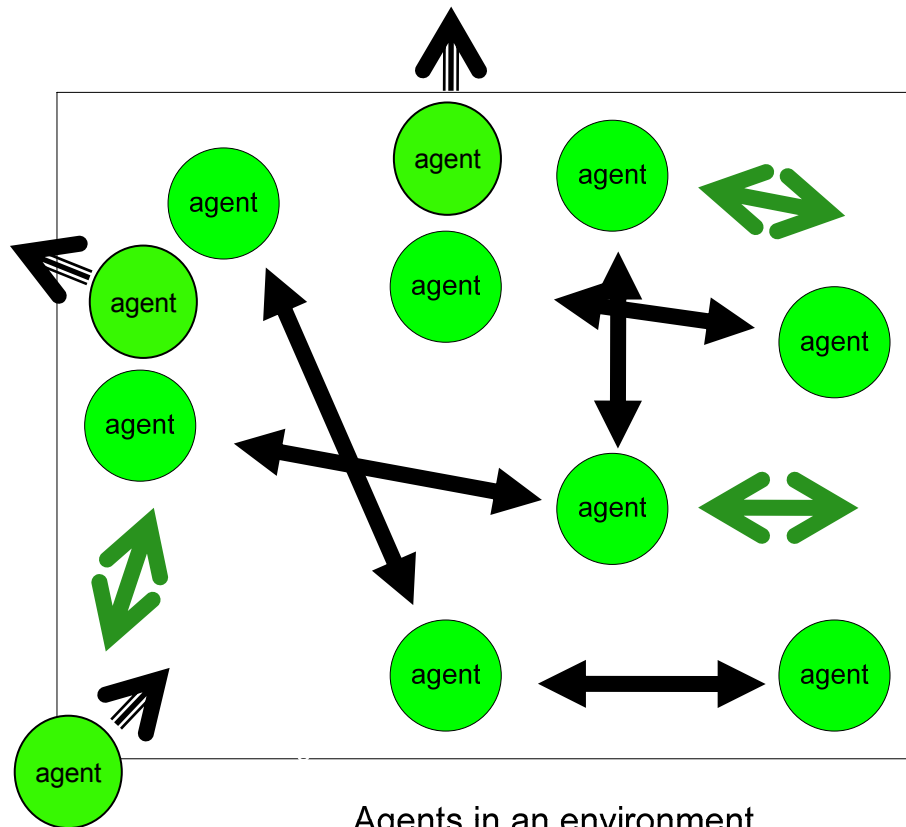


Commonly-used mechanisms in computational simulation

- Multi-agent system as a miniature of human community
- Rule-based system as abstraction of linguistic knowledge
- Iterated communications as abstraction of language contact



Multi-agent system as a miniature of human community





Agent (language user):

- 1) Independent unit;
- 2) Possess certain abilities;

Abilities (linguistic/nonlinguistic behaviors):

- 1) **Memory**: e.g., rule-based system;
- 2) **Activities**: e.g., interaction, replacement;
- 3) Agents are **heterogeneous**;

 Communications among agents
 Interactions with the environment

*Aim: to test whether there are **global tendencies** emerging from the local activities:*

- 1) **Common linguistic knowledge**: mutual understanding;
- 2) **Global social structures**: relationship among agents based on communication;



Rule-based system as abstraction of linguistic knowledge

- Many linguistic phenomena can be viewed as rule-governed (Gumb 1972; Hayes 1989);
- In computational simulation, language exists as a set of linguistic rules among individuals. And the evolution of linguistic knowledge can be viewed as the acquisition and change of linguistic rules in individuals;
- Examples of linguistic rules used in computational simulation:
 - **Lexical rules:** meaning-utterance mappings, “fox” \leftrightarrow /a/ (0.5);
 - Grammatical rules:
 - **Syntactic rules:** word order, SV/VS, SVO/SOV, etc.;
 - **Syntactic categories:** associate lexical rules with syntactic rules so that these lexical rules can be regulated by those syntax rules, S category, V category;

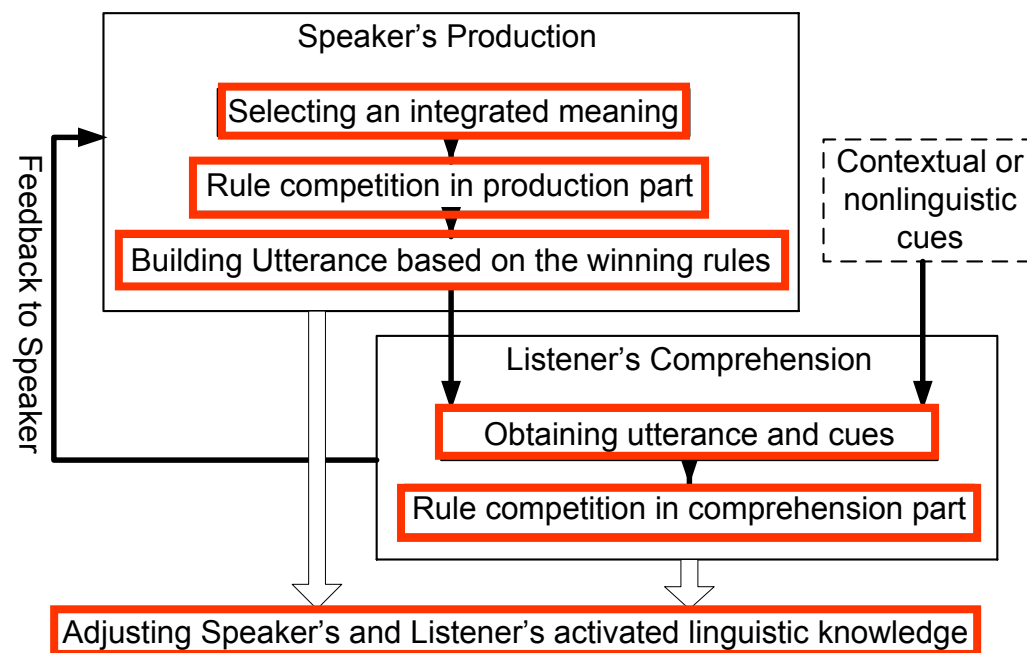
Gumb, R. D. 1972. *Rule-governed linguistic behavior*. The Hague: Mouton.

Hayes, S. C. 1998. *Rule-governed behavior: Cognition, contingencies, and instructional control*. New York: Plenum Press.



Iterated communications as abstraction of language contact

- Language contact is implemented by iterated communications among agents;
- Iterated communications provide opportunities for individuals to exchange linguistic instances and update their linguistic knowledge through the language processing mechanisms in production and comprehension;
- An example of iterated communication:





The compositionality-regularity coevolution model:

Aim: to study the evolution of linguistic universals such as **compositionality** (e.g., lexical items) and **regularities** (e.g., word order) through simulating individual's linguistic behaviors during iterated communications;

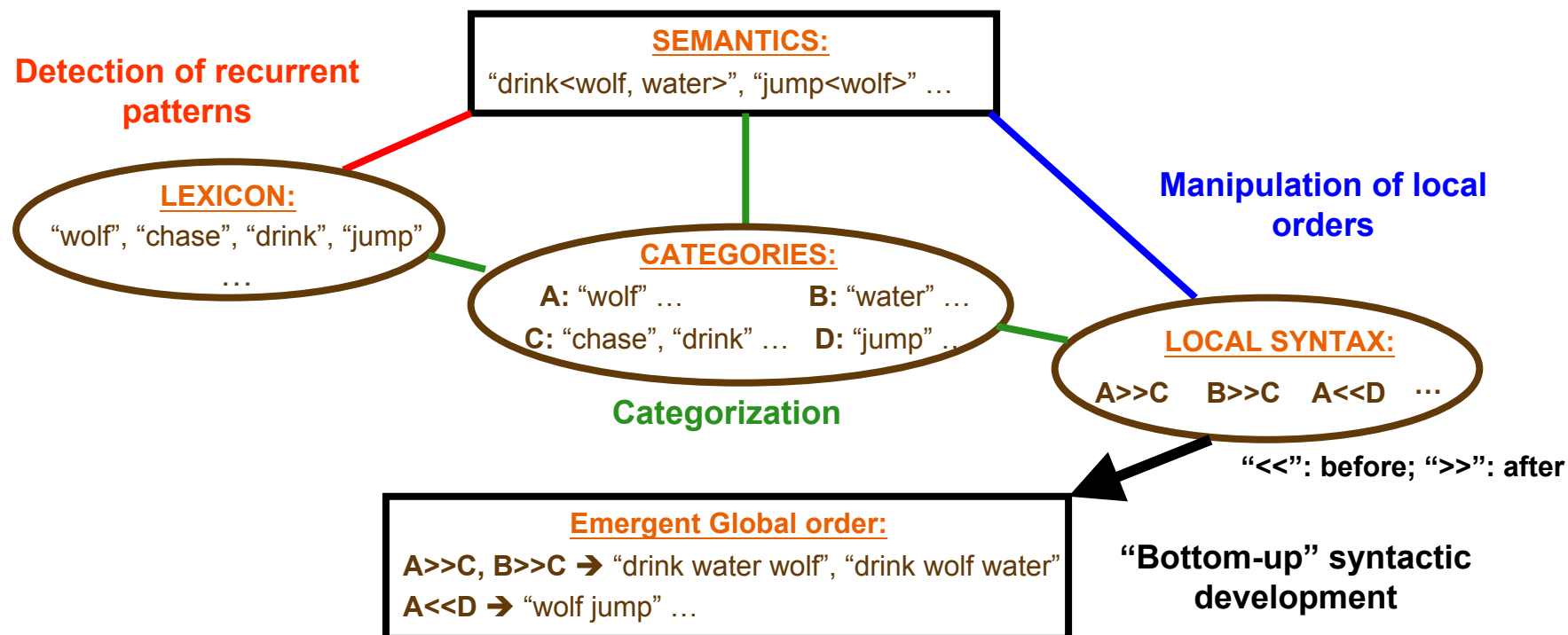
- **Compositionality**: the meanings of complex expressions are determined by the meanings of their components.
 - English: /cats eat rats/;
 - Chinese: /猫吃老鼠/;
 - French: /un chat mange un rat/;
 - Japanese: /猫 が ネズミ を 食べる/;
 - Cat (nominative) rat (accusative) eat (canonical)

- **Regularity at the syntactic level**: many languages adopt conventionalized structures (e.g., word order or morphology) to build up complex expressions.
 - English: */cat rat eats/; /cat eats rat/ and /rat eats cat/;



The conceptual framework

- **Integrated meanings:**
 - Type1: “Pr₁<Ag>”: e.g., “hop<deer>”;
 - Type2: “Pr₂<Ag, Pat>”: e.g., “chase<fox, wolf>”;
- **The conceptual framework:**



Local syntax: binary sequential relation (before or after) between 2 lexical items, e.g., SV, VO;
Global word order (e.g., SVO, SOV) results from local syntax;



The compositionality-regularity coevolution model:

- The major features of this model:
 - Both lexical knowledge and syntactic knowledge are **simultaneously acquired** based on some learning mechanisms;
 - The linguistic processing mechanisms, such as detection of recurrent patterns, categorization, manipulation of local orders are **not language-specific**;
 - The concept of **local order** as binary sequential relation (before or after) between 2 lexical items is introduced, e.g., SV, OV, SO, and **global word order (e.g., SVO, SOV) results from reiterating the available local orders**, e.g., SV+VO→SVO;

- Related papers to this model:
 - Gong, T. 2007. Language evolution from a simulation perspective: On the coevolution of compositionality and regularity. Doctoral Dissertation. The Chinese University of Hong Kong.
 - 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., 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.
 - Ke, J-Y. and Holland, J. H. 2006. Language origin from an emergentist perspective. *Applied Linguistics*, 27(4): 691–716.
 - Gong, T., Minett, J. W., and Wang, W. S-Y. 2006. Computational simulation on the coevolution of compositionality and regularity. In: Cangelosi, A., Smith, A. D. M., and Smith, K., eds., *The Evolution of Language: Proceedings of the 6th International Conference*, London, UK: World Scientific Publishing Co. Pte. Ltd., 99–106.
 - Minett, J. W., Gong, T., and Wang, W. S-Y. 2006. A language emergence model predicts word order bias. In: Cangelosi, A., Smith, A. D. M., and Smith, K., eds., *The Evolution of Language: Proceedings of the 6th International Conference*, London, UK: World Scientific Publishing Co. Pte. Ltd., 206–213.
 - Gong, T., Minett, J. W., and Wang, W. S-Y. 2006. Language origin and the effects of individuals' popularity. *Proceedings of 2006 IEEE World Congress on Computational Intelligence*, Vancouver, CA: 3744–3751.



Two simulation studies on language contact

- The effects of social hierarchy on the maintenance of communal language during intra-group contact
 - Whether certain types of social context can help to maintain an initial compositional language within a group;

- The effects of different linguistic features on the convergence of two communal languages during inter-group contact
 - Given certain degree of inter-group contact, whether two communal languages, by sharing certain linguistic features, can easily converge through sufficient language contact;

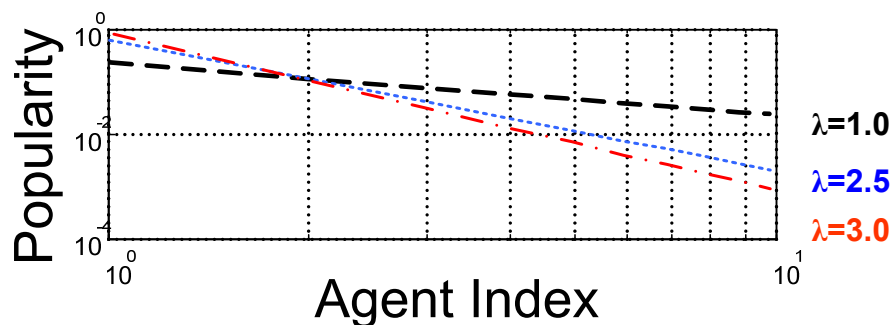


1) The effects of social hierarchy on the maintenance of communal language via intra-group contact

- In human communities, different individuals have different probabilities to communicate with others (**individual's popularity**);
- Different distributions of individuals' popularities reflect social hierarchy in the community;
- **One way to represent these individuals' popularities is Power-law distribution (Wichmann 2005);**

$$y = ax^{-\lambda}$$

$x \rightarrow$ an element or interaction in a given phenomenon,
 $y \rightarrow$ the frequency of this element or interaction.

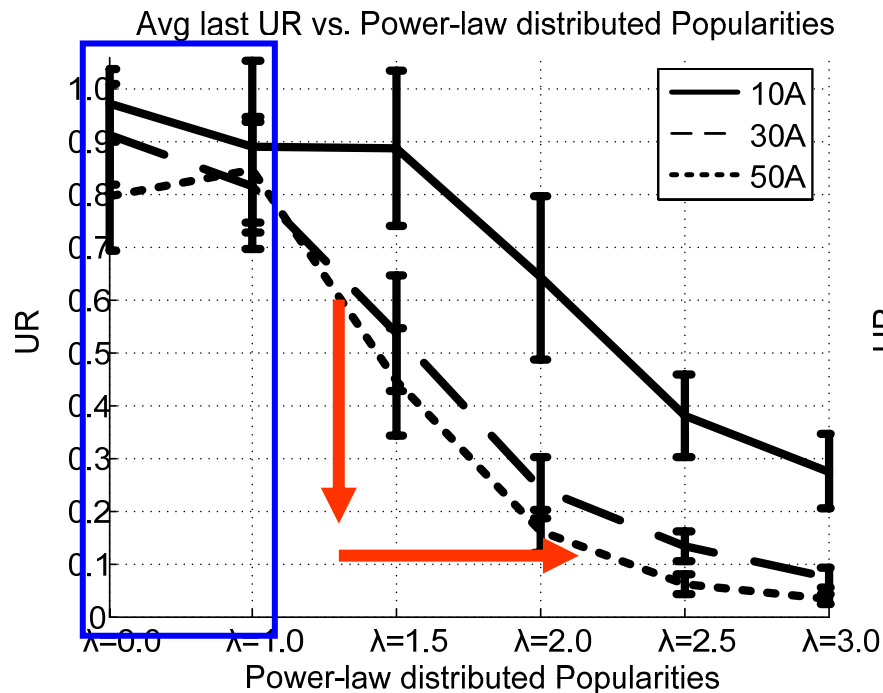


The higher the λ value, the more skewed the social hierarchy

Wichmann, S. 2005. On the power-law distribution of language family sizes. *Journal of Linguistics*, 41(1): 117–131.



1) The effects of social hierarchy on the maintenance of communal language via intra-group contact



Three communities: 10 agents (10A, 6,000 coms), 30 agents (30A, 18,000 coms), 50 agents (50A, 30,000 coms)

Initial stage: all agents share a common compositional language consisting of a set of lexical items and local orders;

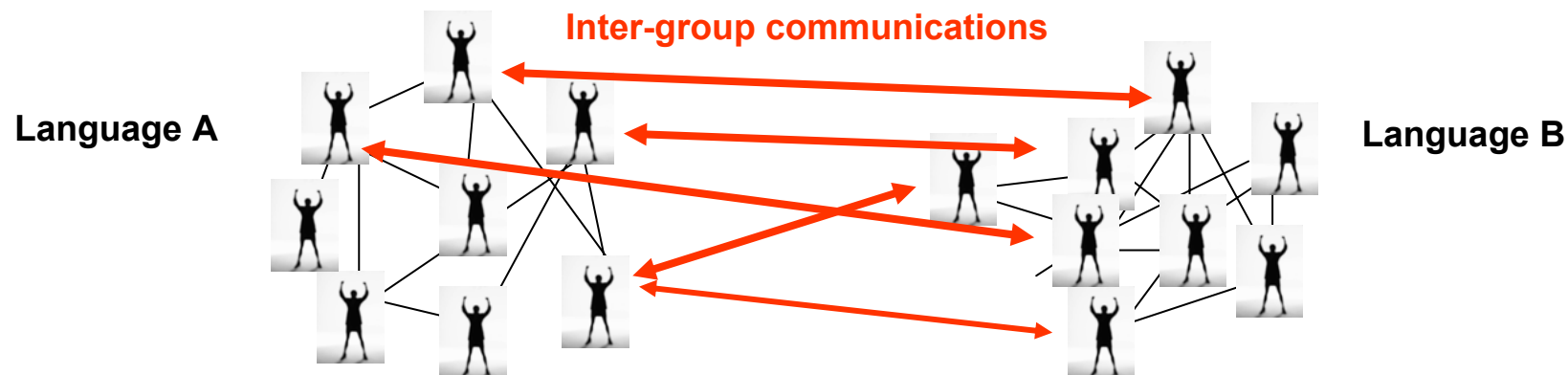
- **UR (Understanding Rate):** the average percentage of integrated meanings that are understandable to all agents in the community; the higher the UR, the higher the probability for the initial language to be maintained;
- In order to maintain a high UR and an initial communal language inside the community, the social hierarchy (λ values) cannot be too skewed (less than 1.0);
- Many language-related social networks that have power-law distributions of language-related interactions do have similar values indicated by this simulation, e.g., email network, phone-call network, etc. (see review in [Newman 2003](#));
- Whys should there exist such boundary λ value (1.0)? **Optimization from two aspects:** a) nonlinguistic aspect; and b) linguistic aspect;

Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM Review*, 45(2): 167–256.



2) The effects of different linguistic features on the convergence of two communal languages via inter-group contact

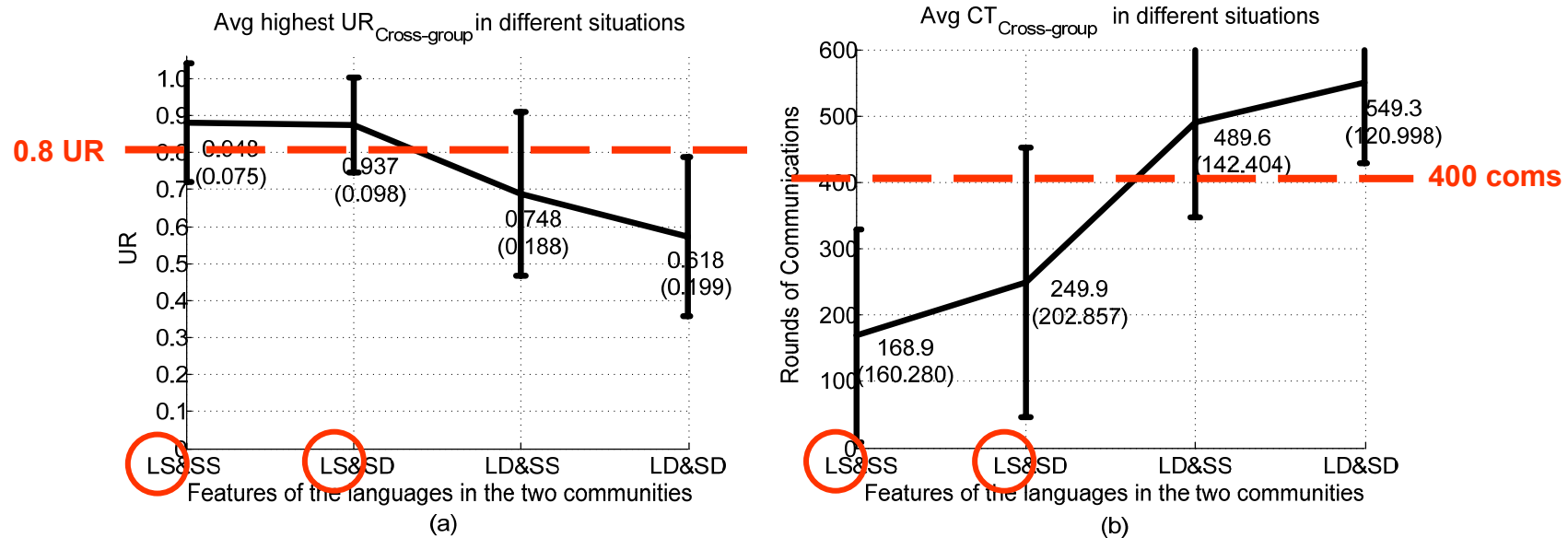
- Linguistic features such as **lexical items and syntax (e.g., word order)** may affect language contact;



- Four situations:**
 - LS&SS:** languages A and B have similar (80%) lexical items and identical word orders: e.g., most dialects within a language;
 - LS&SD:** languages A and B have similar lexical items but different word orders: e.g., Cantonese and Mandarin;
 - LD&SS:** languages A and B have different lexical items but identical word orders: e.g., standard Hindi and standard Urdu (Bhatia 1987);
 - LD&SD:** languages A and B have different lexical items and different word orders: e.g., Tibetan and Mandarin;



2) The effects of different linguistic features on the convergence of two communal languages via inter-group contact



Simulation condition: 20 agents (2 communities), 6000 communications; The percentage of inter-community communication is 60%.

- $UR_{Cross-group}$ → the similarity of two communal languages;
- $CT_{Cross-group}$ → the efficiency of language convergence;
- Sharing similar lexical items is more efficient to converge the two languages than sharing similar word orders;
- Lexical basis of syntax (MacDonald 1994; Bates and Goodman 1997): without similar lexical items, identical word order cannot efficiently help to establish mutual understanding during language contact;

MacDonald, M. C. 1994. Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, 9(2): 157–201.

Bates, E. and Goodman, J. C. 1997. On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes*, 12(5–6): 507–584.



Conclusions and final remarks

- I start from **an empirical study of Daohua**, which reveals that social context and linguistic features play important roles in language contact;
- Apart from the empirical studies, I introduce **computational simulation** to study the influence of these factors on intra- and inter-group language contact;
- I discuss two simulation studies that explore the effects of **social hierarchy** on language maintenance (intra-group), and examine **the influence of lexical items and syntactic features (word order)** on the convergence of different communal languages (inter-group).



Conclusions and final remarks

- Computational simulation is an efficient method to study linguistic problems (Cangelosi and Parisi 2002; Christiansen and Kirby 2003; Cangelosi et al. 2006);
- Computational simulation and empirical studies can assist each other to provide more comprehensive understandings on human language;
- Topics in evolutionary linguistic that can be studied using computational simulation:
 - **Language emergence:** phylogenetic and ontogenetic emergence (Gong, T & Minett, J.);
 - **Language change:** language contact, linguistic innovation, social structure (Gong, T., Minett, J.);
 - **Language death:** language competition, endangered languages field work (Minett, J.);

Cangelosi, A. and Parisi, D. eds. 2002. *Simulating the evolution of language*. London: Springer-Verlag.

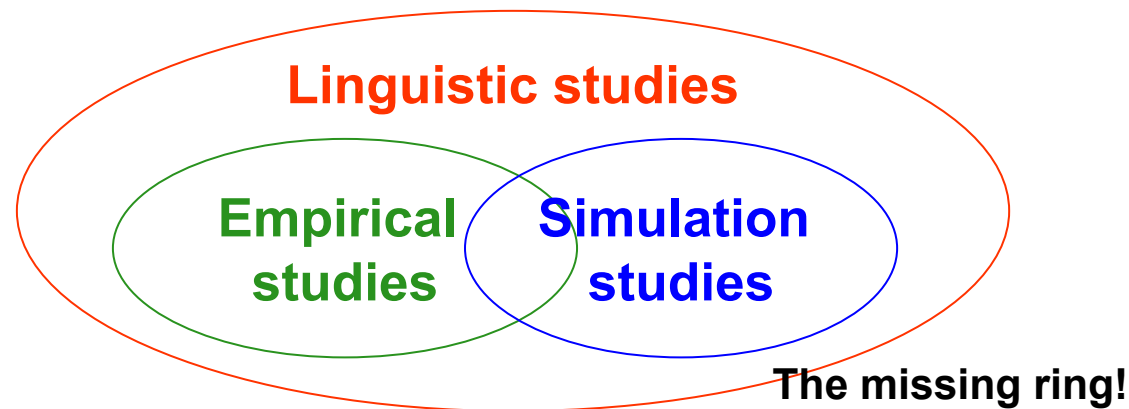
Christiansen, M. H. and Kirby, S. eds. 2003. *Language evolution*. Oxford; New York: Oxford University Press.

Cangelosi, A., Smith, A. D. M., and Smith, K. eds. 2006. *The evolution of language: Proceedings of the 6th international conference*. London: World Scientific Publishing Co.

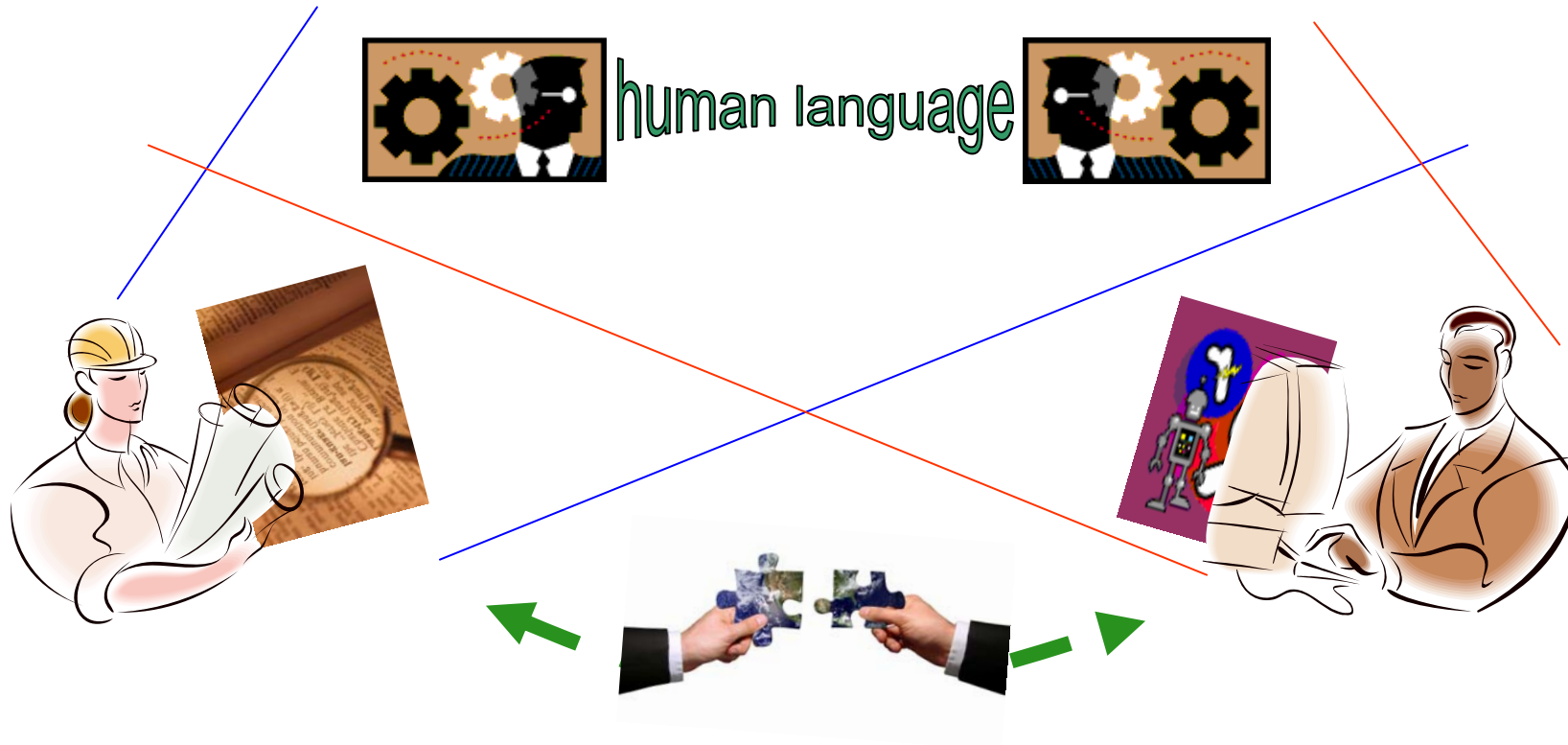


Final remark: *the multi-disciplinary nature of linguistic research*

- Verify and inspire linguistic theories;
- Provide empirical basis for simulated linguistic behaviors and verify simulation results;
- Verify the available linguistic theories and empirical findings;
- Inspire further empirical studies and linguistic theories;



- Insufficient knowledge on artificial intelligence;
- Insufficient technique on complex simulation;
- The complexity of linguistic behavior;
- The authenticity of artificial simulations;



Thank you very much!

Your comments are welcome!