Exploring the Roles of Complex Networks in Linguistic Categorization

Tao Gong*,**
University of Hong Kong

Andrea Baronchelli†
Universitat Politècnica de Catalunya

Andrea Puglisi‡
"Sapienza" Università di Roma

Vittorio Loreto§
Institute for Scientific Interchange
"Sapienza" Università di Roma

Abstract This article adopts the category game model, which simulates the origins and evolution of linguistic categories in a group of artificial agents, to evaluate the effect of social structure on linguistic categorization. Based on the simulation results in a number of typical networks, we examine the isolating and collective effects of some structural features, including average degree, shortcuts, and level of centrality, on the categorization process. This study extends the previous simulations mainly on lexical evolution, and illustrates a general framework to systematically explore the effect of social structure on language evolution.

Keywords
Computer simulation, category game, complex networks

A version of this paper with color figures is available online at http://dx.doi.org/10.1162/artl_a_00051. Subscription required.

1 Introduction

The effect of complex networks on language evolution has been broadly studied in the fields of statistical physics and artificial life using a language-game approach (e.g., [3–7, 9, 12, 17]). This approach views language as a *complex adaptive system* [20, 26], and postulates a population of agents and an interaction protocol (a *language game* [21]) among them. According to this protocol, agents carry out communicative tasks and establish a communication system from scratch, by creating new conceptualization or expression and adjusting available knowledge based on its utility, frequency, or social prestige. After a number of language games, some linguistic knowledge resembling that in real languages is gradually shared among agents. To explore the roles of complex networks, studies often adopt various types of network structures to restrict participants of language games, and use statistical analysis to reveal quantitatively the roles of these constraints in diffusion of linguistic knowledge within or across populations. This line of research sets up a theoretical foundation for the study of language evolution, and the quantitative understanding obtained by statistical analysis can yield important insights into related empirical studies in sociolinguistics [12].

* Contact author.
** Department of Linguistics, University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: gtojty@gmail.com
† Department de Fisica i Enginyeria Nuclear, Universitat Politècnica de Catalunya, Campus Nord B4, 08034 Barcelona, Spain. E-mail: andrea.baronchelli@upc.edu
‡ Dipartimento di Fisica, "Sapienza" Università di Roma, Piazzale Aldo Moro 5, 00185, Roma, Italy. E-mail: Andrea.Puglisi@roma1.infn.it
§ Institute for Scientific Interchange (ISI), Viale Settimio Severo 65, 10133 Torino, Italy; Dipartimento di Fisica, "Sapienza" Università di Roma, Piazzale Aldo Moro 5, Roma, Italy. E-mail: vittorio.loreto@roma1.infn.it

Many contemporary simulation studies that explore the effect of social structure on language evolution bear several limitations. On the one hand, most of the studies (e.g., [5, 9, 12]) focus on lexical evolution, in which language is encoded as form-meaning associations, and language games are simplified as conventionalization of associations among agents. The emergent communication systems become similar to the signaling system [11, 19] and comparable to animal alarm calls that consist of a limited number of concepts and forms. Some language games have been proposed to examine the evolution of linguistic components other than lexical items, such as the guessing game [24] that traces the origin of compositionality in conceptualization and expression, and the *fluid construction grammar* [22] that explores the origins of grammatical constructions to describe tempo-spatial concepts or series of events. However, due to their complexity and focus on language processing, these games either involve only two agents or proceed in a small-scale system, both of which diminish the effect of social factors.

On the other hand, seeking a common lexicon is similar to coordinating actions across agents. Following this perspective, many scholars have developed mathematical models derived from game theory, such as the stag-hunt game [19, 23], to analyze the effect of social structure on coordinative behaviors such as language. In these models, language games are reduced to adjusting one's actions in response to another's, leaving out both the semantic information in linguistic expressions and the necessary processing of linguistic materials.

Language is more than lexical items; semantic and syntactic aspects of language have long been the foci of linguistic research. Likewise, language processing and communication are more than action coordination; categorical or syntactic operations are not solely for coordination, and they may introduce ambiguity in communications. Moreover, social constraints may cast their influence on the evolution of linguistic components other than lexical items, and such influence may manifest itself in different ways. For instance, during language contact, the same sociocultural setting may lead to distinct borrowing strategies with respect to lexical and grammatical items among competing languages (e.g., [1]). All these indicate that before a clear analysis on the processing of various linguistic components, it is presumptuous to generalize the conclusions drawn from those game theory models to other aspects of language evolution.

A theoretical study exploring the roles of complex networks in language evolution should proceed as follows:

(i) Design a language game involving a particular linguistic component (e.g., lexicon).

(ii) Analyze the dynamics of this game first in a system without social constraints, and then in a system with social constraints.

(iii) Extend the current game to consider other linguistic components, and repeat (ii).

According to this framework, lexical evolution models only perform steps (i) and (ii), and game theory models simply skip step (iii) to reach ultimate conclusions. Therefore, all these models remain incomplete. Only by repeating these steps can we approach a comprehensive understanding of the effect of social structure on language evolution.

As an example of fulfilling step (iii), we adopt the category game model [17], which extends the lexical evolution model of the naming game [3], and apply it in different networks to study the effect of social structure on linguistic categorization. Besides lexical items, the category game simulates the origins and evolution of categories to discriminate stimuli from a perceptual channel. The categorical operations include segmenting perceptual space into perceptual categories and coordinating their lexical items. In the rest of this article, we describe the category game and analyze its dynamics in a fully connected network (Section 2), put this game in typical networks to discuss the effects of social factors on the categorization process (Section 3), and finally, conclude the article (Section 4).

2 The Category Game and Its Dynamics in a System without Social Constraints

The category game model was designed to simulate the origins of color categories in human languages. The basic model involves a population of N artificial agents. Starting from scratch, these agents can

automatically generate, via a number of *category games*, a categorization pattern for the visible light spectrum shared in the whole population. For the sake of simplicity and without loss of generality, color perception is reduced to a single analogical continuous perceptual channel, and each stimulus becomes a real number from the interval [0,1) that represents its normalized, rescaled wavelength. *A categorization pattern* is a partition of the interval [0,1) into subintervals, or *perceptual categories*. Each agent has its own word inventories that link perceptual categories with their linguistic counterparts (i.e., lexical items), and these inventories evolve during iterated games among agents.

A category game between two agents (a speaker and a hearer) proceeds as follows. First, a scene of $M \geq 2$ stimuli is presented to these agents; no two of the stimuli can appear at a distance smaller than a parameter *dmin* (similar to the just-noticeable difference in psychophysics), and one of them becomes the *topic* of this game. Then, the speaker discriminates the scene, if necessary refines its perceptual categories via categorization mechanisms, and utters the lexical item that is associated with the topic. After that, the hearer tries to guess the topic, based on its own categorization knowledge and the heard word. According to the outcome (success or failure) of the game, both agents align their form-meaning inventories, by inventing new categories or adjusting the lexical terms of available ones. Two examples of the category game are shown in Figure 1. Further technical details of the model can be found in [4, 17].

This model has the advantage of incorporating an extremely small number of parameters—basically, the number of agents (N) and the discrimination curve *dmin*—but generating rich and realistic output.

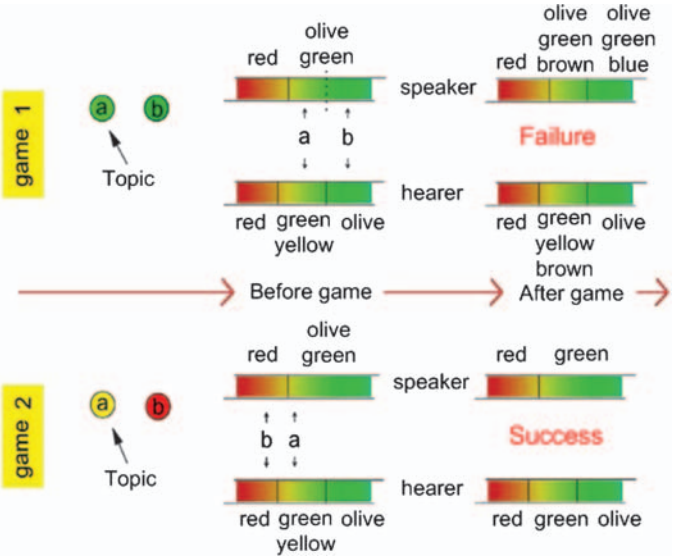

Figure 1. Two examples of the category game (adapted from [17]). Balls denote stimuli presented in these games, among which the topics are identified by arrows. Colorful banners denote perceptual channels, and different agents use different bars to partition the channels into perceptual categories, whose word inventories are listed above or below. In game 1, the two stimuli fall into the same perceptual category in the speaker; then, the speaker discriminates the topic ("*a*") by creating a new boundary in this category at the position $(a + b)/2$. In this way, two new categories are created, both inheriting the word inventory ("green" and "olive") of their parent category, and two new words are invented, respectively, in these new categories ("brown" and "blue"). After that, the speaker browses the list of words associated with the category that contains the topic. There are two cases here: If a previous successful game occurred using this category, its last winning word is chosen; otherwise, the newly created word ("brown") is chosen, and sent to the hearer. Since the hearer does not have this word in its inventory, this game fails. Then, the speaker clarifies the topic by pointing, and the hearer discriminates the topic and adds the word "brown" to the inventory of the corresponding category that can discriminate the topic. In game 2, the topic "*a*" is discriminated by the perceptual category whose last winning word is "green." Then, the speaker sends "green" to the hearer. The hearer knows this word, and the perceptual category containing this word can discriminate the topic. If ambiguity arises in the sense that the speaker's word is associated with more than one category that contains the topic, an unbiased random choice is taken by the hearer. In this example, the game succeeds. Then, similarly to the naming game, both agents try to align their categories, by eliminating competing words in their used categories and leaving the used word "green" only. This alignment strategy adjusts word inventories of categories, not boundaries of categories.

We first analyze the dynamics of the category game in a fully connected network without social constraints. In this condition, each game occurs between two randomly chosen agents in the population. The evolution of categories proceeds in the following two phases.

1. $0–10^4$ games per agent: Agents initially have one perceptual category covering the whole perceptual channel, and no words. The pressure for discrimination causes the number of perceptual categories to increase, and many new words are associated with categories among agents. This kind of synonymy reaches a peak, and then dries out, in a way similar to the naming game. Game 1 in Figure 1 illustrates how perceptual categories are created and how their inventories are expanded.
2. $10^4–10^6$ games per agent: When on average only one word is recognized by the whole population for each perceptual category, the second phase of evolution intervenes. In this phase, words expand their reference across adjacent perceptual categories, merging these categories into *linguistic categories*. The coarsening of categories becomes slower and slower, with a dynamic arrest analogous to the physical transformation of a glass-forming liquid into a glass upon rapid cooling [6]. Two examples in Figure 2 illustrate how adjacent perceptual categories acquire identical word forms and merge into a linguistic category. This phase starts at about 10^4 games per agent (based on the simulation results in different population sizes), during which the linguistic categorization pattern can reach a 90% to 100% degree of sharing among agents and remain stable for a plateau phase. The duration of the stable stage diverges with the population size [17].

If one waits for a much longer time, say 10^7 games per agent, the number of shared linguistic categories starts dropping, which is mainly caused by the extremely slow diffusion of category boundaries, due to the finite-size effects [6]. As examined in [15], the category game on this time scale shows the glassy behavior [14]. That is to say, the relaxation time of the number of shared linguistic categories increases

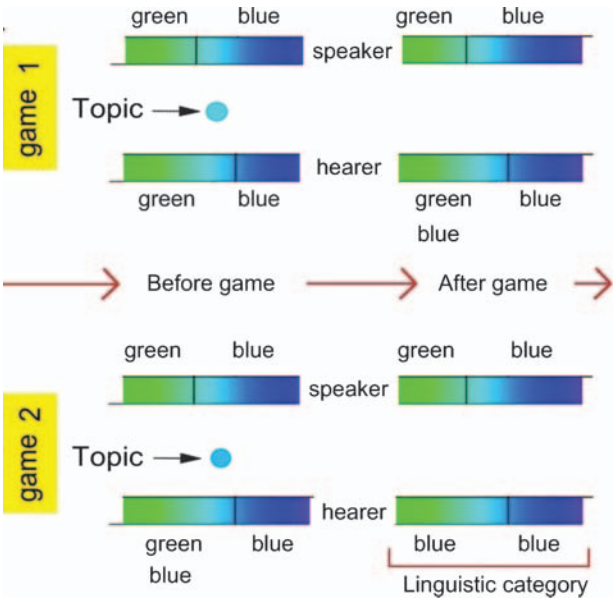

Figure 2. Two examples of the category game in the second phase of evolution (adapted from [17]). In game 1, the speaker describes the topic with the word "blue." In this failed game, the hearer learns "blue" in one of its perceptual categories. In game 2, the speaker describes a similar stimulus using the same word "blue." In this successful game, the hearer removes the competing word "green" and keeps "blue" in that category. Now, the two adjacent perceptual categories in the hearer use the same word "blue," which thus become a linguistic category.

massively, as in a singularity, at a finite number of categories; the system traps itself in a metastable state; and the dynamics is arrested even though the final state (only one category) is not reached. Meanwhile, as proved in [6, 15], the number of linguistic categories is weakly dependent on the population size. Since comparison with the real world is less accessible on such a long time scale, we disregard the glassy behavior of the model on this time scale, but treat the second phase as the stable stage of the model.

As seen from the examples in Figures 1 and 2, individual behaviors include both simple coordinating behaviors as in the naming game (e.g., in successful games, used words are preserved and competing ones are deleted), and categorical operations helping to discriminate stimuli (e.g., the perceptual space is separated into categories, and new lexical terms for stimuli in those categories are induced). The categorical operations, resembling those in processing linguistic categories, may not lead to immediate coordination; the division of the perceptual space is not necessarily aligned. However, after iterated games among agents, a common categorization pattern (the whole perceptual channel is roughly separated into a limited number of linguistic categories, and most agents use identical lexical terms to name stimuli from those categories) will get shared among agents.

Such a two-phase evolution can be traced by several indices. First, the *overlap* (*OL*), measured as in Equation 1, evaluates the alignment of perceptual or linguistic categories. A high *OL* reflects a high degree of coordination on discriminating stimuli from the perceptual space:

$$OL = \frac{2\sum_{i<j} o_{ij}}{N(N-1)}, \qquad o_{ij} = \frac{2\sum_{c_{ij}} (lc_{ij})^2}{\sum_{c_i} (lc_i)^2 + \sum_{c_j} (lc_j)^2} \tag{1}$$

where lc is the width of category c, c_i is a category from player i, and c_{ij} is the category intersection set obtained based on category boundaries of both players i and j. The ratio o_{ij} indicates the degree of alignment between categories of players i and j, which reaches 1.0 if the boundaries of the two sets of categories are identical.

Second, the *understanding rate* (*UR*, similar to the success rate in [17]) calculates the percentage of successful category games between all pairs of agents in the population. A high *UR* reflects a high degree of conventionalization of lexical terms in categories among agents. To measure *UR*, we let agents play *virtual* category games without updating their categories and inventories, and calculate the percentage of successful games in these virtual games.

Apart from these indices, the shrinkage of synonyms can be traced by the average number of words per category, and the coarsening of linguistic categories by the number of shared words in the population. In the second phase of evolution, each shared word usually corresponds to one linguistic category shared among agents.

Based on these indices, Figure 3 traces the evolution of linguistic categories after 5×10^7 games per agent in a fully connected network ($N = 100$, $dmin = 0.01$). As in Figure 3a, low perceptual *OL* (dashed line) and high linguistic *OL* (solid line) indicate that although perceptual categories keep emerging, some of them already merge into linguistic categories, each having similar boundaries across agents. As shown in Figure 3b, with a logarithm axis, *UR* starts from a low value, then undergoes a sharp transition, and finally exceeds 0.9, tracing a transition from no mutual understanding to a common categorization pattern among agents. Figure 3c and d illustrate the two-phase evolution: After 10^4 games per agent, the shared categorization pattern remains stable and *UR* is high; after 10^6 games per agent, the number of shared words (linguistic categories) begins dropping.

3 The Roles of Complex Networks in Linguistic Categorization

We now turn to network structures. Following the approach in the previous work (e.g., [8–10]), we treat agents as nodes and their interactions as edges among nodes. We consider five types of networks: ring,

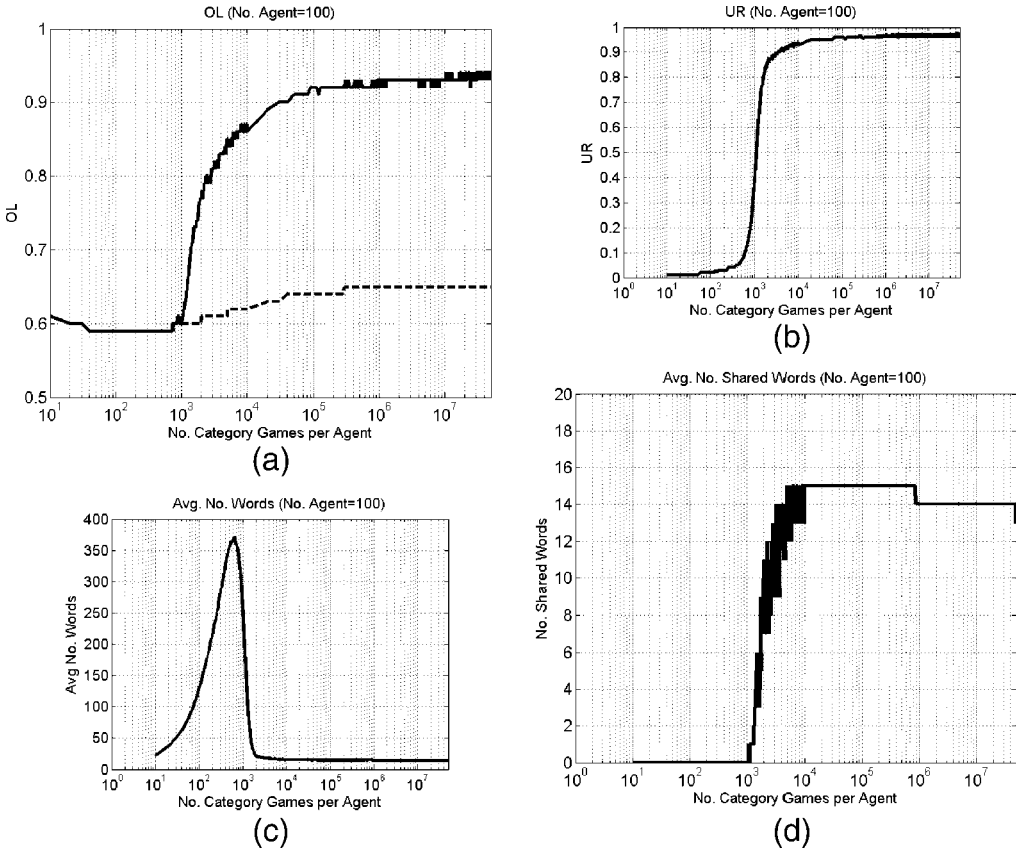

Figure 3. Results of one simulation in a fully connected network ($N = 100$, $dmin = 0.01$): (a) perceptual *OL* (dashed line) and linguistic *OL* (solid line); (b) *UR*; (c) average number of words per category; (d) average number of shared words.

two-dimensional lattice, small-world network [27], scale-free network [2], and star network. These structures characterize several key features of real-world communities [16]. For example, small-scale societies are usually fully connected or have a starlike, centralized structure; social connections among geographically distributed communities can be denoted by a ring or lattice; and large-scale societies usually show small-world or scale-free characteristics. These networks are clarified in Table 1, using

Table 1. Five types of networks and their characteristics (based on 100 nodes). Scale-free network is formed by preferential attachment [27]; each new node has two connections to previous ones, so that the average degree is around 4. Small-world network is formed by rewiring from a 2D lattice [2], with the rewiring rate as 0.1. Numbers within parenthesis are standard deviations of the values in scale-free and small-world networks.

Network type	Average degree (*AD*)	Clustering coefficient	Average shortest path length (*ASPL*)
Ring	2	0.0	25.25
Two-dimensional lattice	4	0.5	12.88
Small-world	4	0.17 (0.031)	3.79 (0.086)
Scale-free	3.94 (4e−14)	0.14 (0.038)	3.01 (0.071)
Star	1.98	0.0	1.98

indices such as the *average degree* (*AD*, the average number of edges per node), *clustering coefficient* (the probability for neighbors of a node to be neighbors as well), and *average shortest path length* (*ASPL*, the average smallest number of edges via which any two nodes can connect).

Instead of immediately focusing on certain indices and manipulating a particular type of network according to these indices (e.g., [19, 23]), we first browse the general performance of the category game in these networks to reveal social factors that could potentially lead to differences in performance, and then analyze those factors in particular via additional manipulation. To evaluate the generality of the results, we consider populations with different sizes, including 100, 200, 500, 800, and 1000, among which population 100 is the focus. In each simulation, 10^6 games per agent are conducted. In a network, only directly connected agents can play category games. We conduct 20 simulations in each type of network and measure the average linguistic *OL*, *UR*, and the number of shared words in these simulations.

Figure 4a–c show the simulation results in a 100-agent population. The results in a fully connected network are also included for comparison. We conduct a one-way analysis of covariance (ANCOVA) (dependent variable: *UR* in 20 simulations; fixed factor: Six types of networks, including a fully connected network; covariate: 201 sampling points along 10^6 games per agent, in which 101 points lie in $0–10^4$, with a step of 100, and 100 points lie in $10^4–10^6$, with a step of 10^4) to evaluate the effect of network structures on *UR*. This analysis shows that network structures have a significant effect on

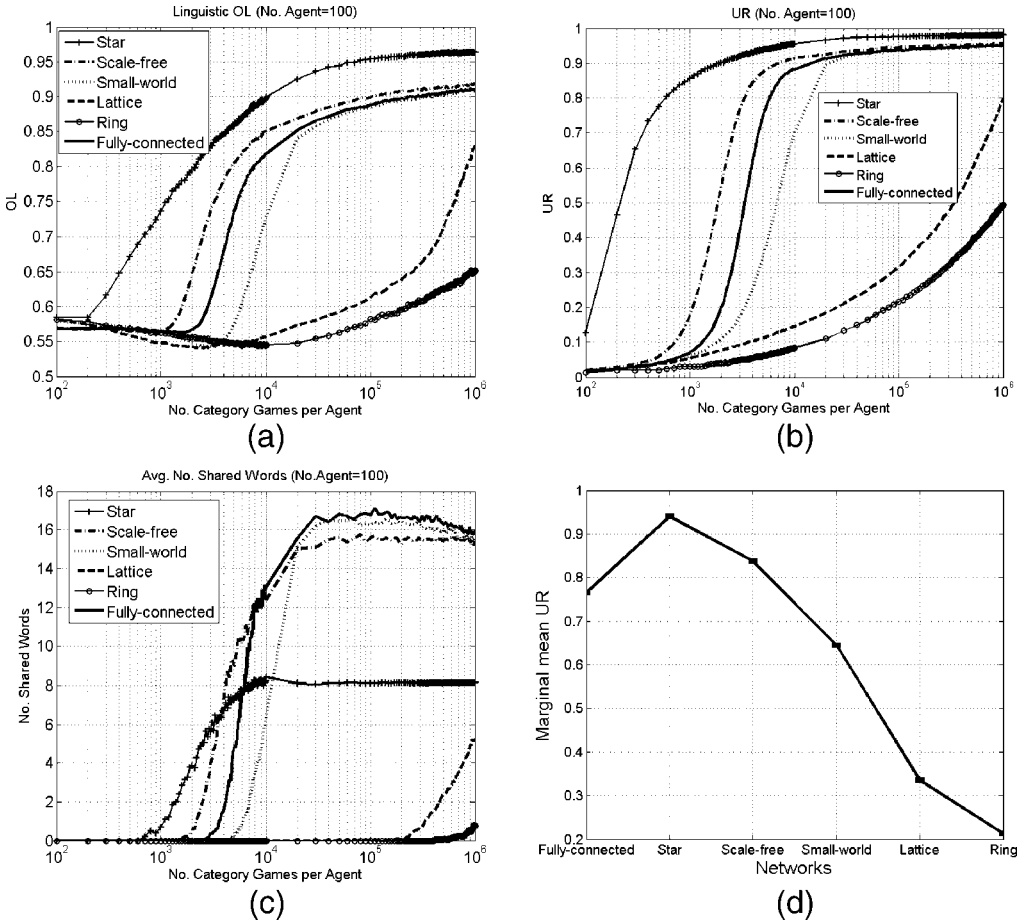

Figure 4. Results in different types of network ($N = 100$, $dmin = 0.01$): (a) linguistic *OL*; (b) *UR*; (c) average number of shared words; (d) marginal mean *UR* in different networks. Each line in (a)–(c) is averaged over 20 simulations in the same condition.

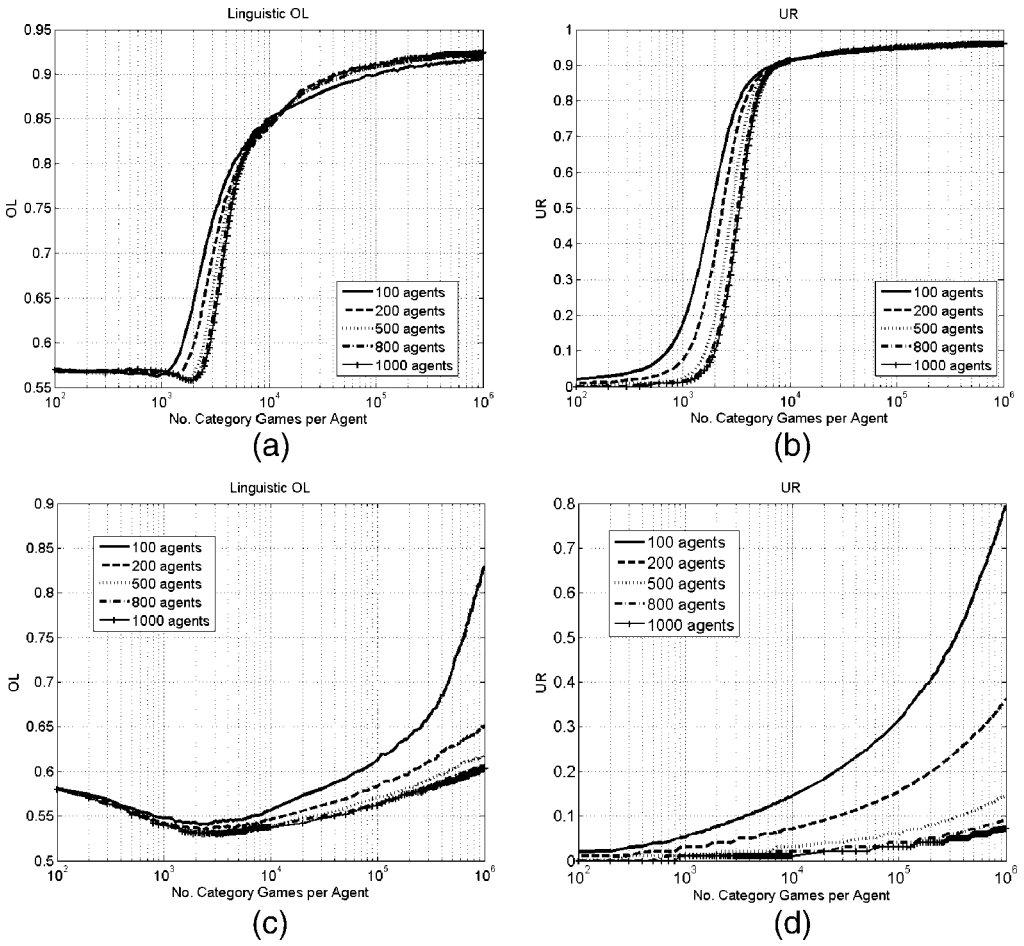

Figure 5. Results in populations of different sizes: Linguistic *OL* in scale-free network (a) and lattice (c), and *UR* in scale-free network (b) and lattice (d). Each line is averaged over 20 simulations in the same condition.

UR ($F(5, 23993) = 334.885$, $p < 0.001$, $\eta_p^2 = 0.657$), as shown in Figure 4d (the sampling points also have a significant effect, $F(1, 23993) = 560.543$, $p < 0.001$, $\eta_p^2 = 0.391$).

Compared with the fully connected network, the categorization process is accelerated in scale-free and star networks, but *delayed* (the sharp increase in linguistic *OL* occurs much later, and so does the sharp increase in *UR*) in small world, lattice, and ring. This tendency also manifests in the marginal mean *UR*. Due to the delay, the average number of shared words in ring or lattice is much lower than those in the other networks. Although the star network has about as high a *UR* as the scale-free, small-world, and fully-connected networks, the number of shared words in the star network is less than half of those in the other networks. The number of shared words in the scale-free network is also lower than those in the small-world and fully connected networks.

Figure 5 compares the results in populations having 200, 500, 800, and 1000 agents. This comparison focuses on two aspects. The first is the pace of evolution. Figures 5a and b show the linguistic *OL* and *UR* in the scale-free network, and Figures 5c and d show these indices in the lattice. In the scale-free network, with the increase in the population size, the emergence of a common categorization pattern across agents is delayed, but after some number of games, the values of those indices remain similarly high. In the lattice, however, with the increase in the population size, the emergence process occurs much later, and different populations have different linguistic *OL* and *UR*. In addition, in an 800- or

1000-agent lattice, after 10^6 games, both linguistic *OL* and *UR* are low, indicating a low degree of sharing a common categorization pattern.

The second aspect of comparison concerns the values of indices across populations. Figure 6 compares the linguistic *OL*, *UR*, and the average number of shared words in different networks. As shown in Figures 6a and b, with the increase in the population size, linguistic *OL* and *UR* are similarly high in the star, scale-free, and small-world networks, but are much lower in the lattice and ring. As shown in Figure 6c, with the increase in the population size, the number of shared words is similarly high in the scale-free and small-world networks, but about half of the value in the star network and zero in the lattice and ring.

These figures show that linguistic categorization has different dynamics in different networks, and these dynamics are less dependent on the population size. Combining Table 1 with Figure 4, we observe a correlation between linguistic *OL* (and *UR*) and *ASPL* (average shortest path length): With the decrease in *ASPL* from ring to lattice, small-world, scale-free, and star networks, the emergent linguistic categories show higher *OL* and *UR*, and the sharp increases in these indices occur much earlier.

A small *ASPL* can be achieved in three ways. The first is to increase *AD* (average degree), so that nodes can connect directly with many others. This reduces the number of intermediate nodes for connecting any two nodes in the network, thus decreasing *ASPL*. As seen in Figure 4, from ring to lattice, *AD* increases from 2 to 4, and the emergence of shared linguistic categories is accelerated.

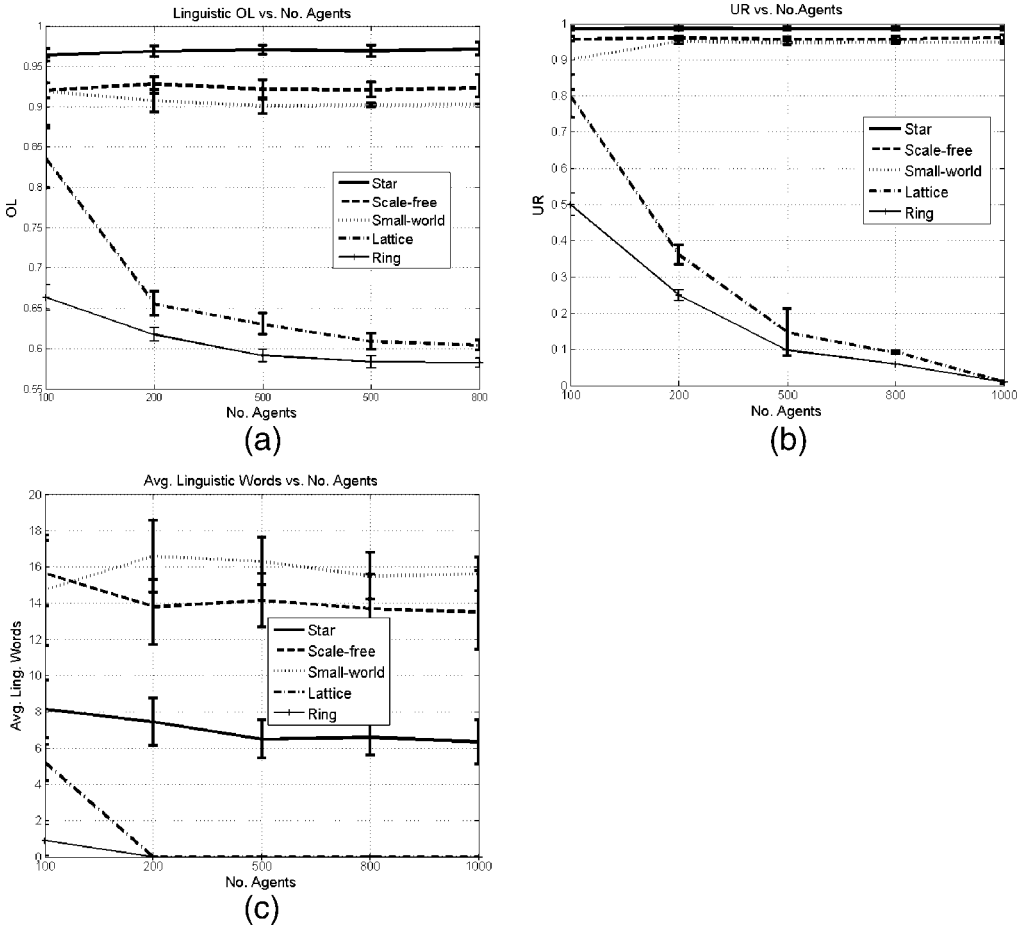

Figure 6. Results in populations with different sizes: (a) linguistic OL; (b) UR; (c) average number of shared words. Each line is averaged over 20 simulations in the same condition, and error bars denote standard deviations.

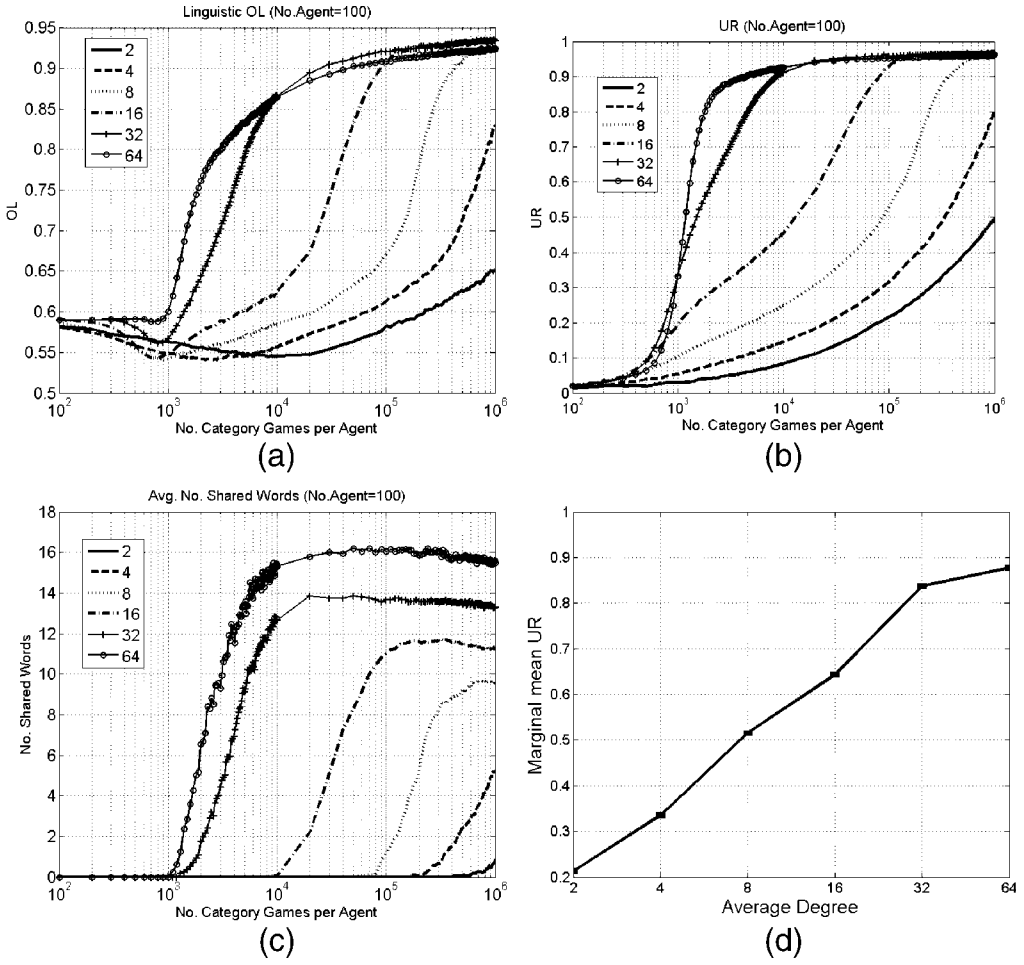

Figure 7. Results in lattices with different *AD* ($N = 100$, $dmin = 0.01$): (a) linguistic *OL*; (b) *UR*; (c) average number of shared words; (d) marginal mean *UR*. Each line in (a)–(c) is averaged over 20 simulations with a particular *AD*.

We further set up simulations in lattices with *AD* as 2 (ring), 4 (lattice), 8, 16, 32, and 64, and conduct an ANCOVA (dependent variable: *UR* in 20 simulations; fixed factor: *AD*; covariate: 201 sampling points throughout 10^6 games per agent) to evaluate the effect of *AD* on *UR*. As shown in Figure 7a–c, with the increase in *AD*, the increase in linguistic *OL* occurs earlier, so does the increase in *UR*, and there are more shared words. The ANCOVA analysis shows a significant effect of *AD* on *UR* ($F(5, 23993) = 285.759$, $p < 0.001$, $\eta_p^2 = 0.652$) (the sampling points also have a significant effect: $F(5, 23993) = 877.754$, $p < 0.001$, $\eta_p^2 = 0.536$), as shown in Figure 7d. These results confirm that *AD* can affect linguistic categorization.

The second way of reducing *ASPL* is to introduce direct connections (or *shortcuts*) to nodes that are not locally connected in a network. Via shortcuts, many agents can directly affect each other, and connections between any two nodes in a network are shortened. This explains the difference in performance between the lattice and the small-world network, of which both have the same *AD* (=4), but the latter has some shortcuts.

We further evaluate the effect of shortcuts on linguistic categorization, by adjusting the rewiring rate from 0.01 to 0.02, 0.05, 0.08, and 0.1. These rates preserve the small-world characteristic in the network, and with the increase in rewiring rate, the proportion of shortcuts increases as well. We also conducted an ANCOVA (dependent variable: *UR* in 20 simulations; fixed factor: rewiring

rates; covariate: 201 sampling points throughout 10^6 games per agent) to evaluate the effect of shortcuts on *UR*. As in Figure 8a–c, with the increase in rewiring rate, the increase in linguistic *OL* occurs earlier, so does the increase in *UR*, and there are more shared words. The ANCOVA analysis shows a significant effect of shortcuts on *UR* ($F(4, 19994) = 673.611$, $p < 0.001$, $\eta_p^2 = 0.119$) (the sampling points also have a significant effect: $F(4, 19994) = 33667.205$, $p < 0.001$, $\eta_p^2 = 0.627$), as shown in Figure 8d. These results confirm that shortcuts can affect linguistic categorization.

The third way of reducing *ASPL* is to organize the networks in such a way that many nodes are directly connected to some particular nodes (*hubs*). These hubs serve as the intermediate nodes to connect other nodes. Such a centralized structure gives hubs more chances to participate in games, develop their categorization patterns, and affect others. All these aspects help accelerate the conventionalization of linguistic categories.

Centrality explains the performances in the star and scale-free networks. Via preferential attachment, the scale-free network possesses many hubs. In the star network, there is a supernode connecting all other nodes. Such an extremely centralized structure leads to high linguistic *OL* and *UR*, and an early increase in these indices. Centrality also affects *AD*: It may increase *AD*, as in the scale-free network, or decrease *AD*, as in the star network. Although it is difficult to isolate centrality from

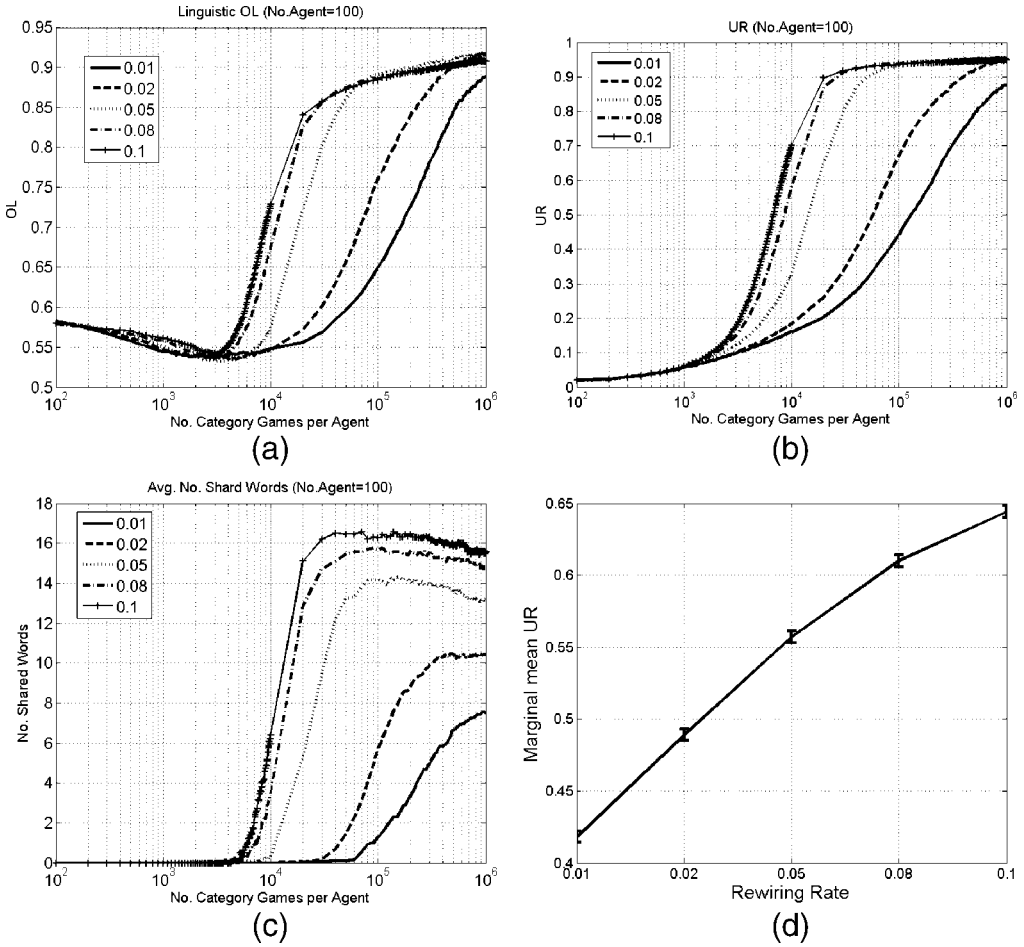

Figure 8. Results in small-world networks with different rewiring rates ($N = 100$, $dmin = 0.01$): (a) linguistic *OL*; (b) *UR*; (c) average number of shared words; (d) marginal mean *UR*. Each line in (a)–(c) is averaged over 20 simulations with a particular rewiring rate.

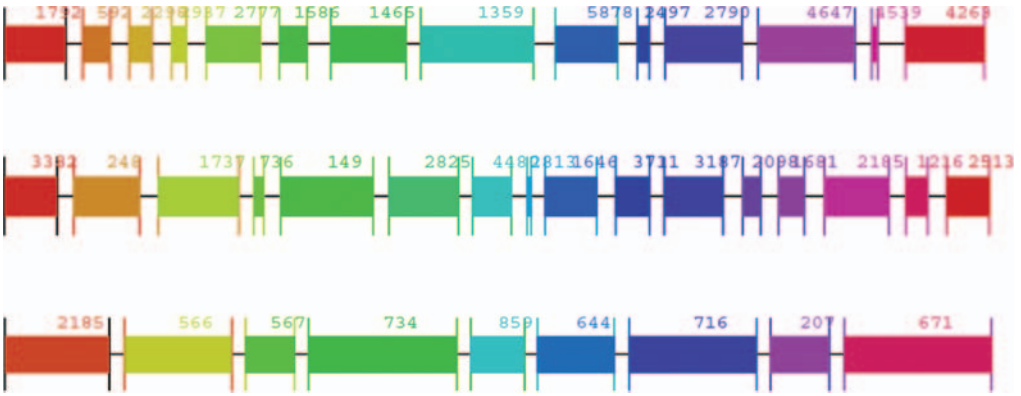

Figure 9. Emergent common linguistic categories in small-world (top panel), scale-free (middle panel), and star (bottom panel) networks. In each panel, the horizontal line denotes the perceptual channel [0,1), the vertical bars mark the boundaries of linguistic categories (colored patches) shared by all agents in the population, and the word forms of these categories are listed above. Uncolored regions between bars and outside colored patches are uncertain regions.

other features such as *AD*, the comparison of the results in the star and scale-free networks indicates that the level of centrality is an important factor affecting linguistic categorization.

These structural features (*AD*, shortcuts, and level of centrality) collectively lead to the results in Figure 4. The ring has the lowest *AD*, no shortcuts, and the lowest level of centrality among these networks. After a fixed number of games, both the linguistic *OL* and the *UR* in this network remain the lowest. Due to the increase in *AD*, the performance in the lattice is improved. Due to the involvement of shortcuts, the performance in the small-world network is further improved. Due to the level of centrality, the performance in the scale-free network becomes much better. And finally, although *AD* in the star network is the smallest, its high level of centrality makes its linguistic *OL* and *UR* the highest among all networks.

The above discussions are based on linguistic *OL* and *UR*. For the average number of shared words, the star, scale-free, and small-world networks also show different performance: Despite similarly high linguistic *OL* and *UR*, the number of shared words in the star network is less than half of those in the scale-free and small-world networks.

These performance results can be ascribed to the level of centrality. First, the extremely centralized structure in the star network allows the supernode to participate in all games with others, which causes the supernode to quickly develop its linguistic categories and affect others' categorization patterns. The evolution of linguistic categories in this network occurs in an intensively cohesive social environment. During the second phase of evolution, words can easily expand their references across adjacent perceptual categories, and those categories can easily merge into linguistic categories. Then, the number of shared linguistic categories in this network remains small. Second, the scale-free network is less cohesive than the star network, and some perceptual categories may not be easily merged into linguistic categories, leaving many subintervals in the perceptual channel. Then, the number of shared linguistic categories in the scale-free network is higher than that in the star network. Third, shortcuts in the

Table 2. Widths of uncertain regions and *UR*. Values within parenthesis are standard deviations.

Network	Average width of uncertain regions	Average *UR*
Small-world	0.2716 (0.0352)	0.9495 (0.0088)
Scale-free	0.2455 (0.0307)	0.9555 (0.0076)
Star	0.1012 (0.0228)	0.981 (0.0064)

small-world network are less efficient than hubs in the scale-free network for increasing the cohesion of the network. Therefore, the number of shared linguistic categories in the small-world network is higher than that in the scale-free network.

These predictions can be verified by *uncertain regions* in the perceptual channel. For stimuli from these regions, agents do not describe color by means of a uniform lexical term. Figure 9 shows examples of the shared linguistic categories that emerge respectively in the small-world, scale-free, and star networks with 100 nodes. In these perceptual channels, uncertain regions refer to those uncolored parts. Table 2 records the widths of uncertain regions and *UR* after 10^6 games per agent in these networks.

The width of uncertain regions reflects the difficulty in merging adjacent perceptual categories into linguistic categories. As seen from Table 2, the width of uncertain regions is the biggest in the small-world network, but the smallest in the star network. As shown in Figure 9, some categorized regions in the perceptual channel that emerge in the small-world network are narrow and surrounded by wide uncertain regions. It is difficult for such categories to merge. This reflects the less cohesive nature of the small-world network and its bigger number of shared categories than in the scale-free and start networks. Also, although *UR* keeps increasing with the decrease in the width of uncertain regions, the similarly high *UR* in these networks indicates that these categorization patterns are sufficient to help discriminate stimuli from many regions of the perceptual channel.

These conclusions also hold in bigger populations. Table 3 lists the widths of uncertain regions and *UR* in populations with 100, 200, 500, 800, and 1000 agents under the small-world, scale-free, and star networks. With the increase in the population size, the width of the uncertain regions in the first two types of networks (but not the star network) increases, but *UR* remains high.

4 Discussion and Conclusions

This article discusses the roles of complex networks in linguistic categorization based on a category game model that simulates the coevolution of linguistic categories and their lexical terms. According to the results in some typical networks, we identify three social factors—namely average degree, shortcuts, and level of centrality—that can affect linguistic categorization in terms of alignment, understandability, and number of shared linguistic categories. Increasing the average degree or level of centrality can accelerate the categorization process; introducing shortcuts can serve this purpose to a certain extent; and increasing the level of centrality can reduce the number of shared linguistic categories. These findings are not greatly affected by the population size, and consistent with the conclusions of a recent empirical study based on real languages [13].

Some of these conclusions are in line with those drawn from game-theory-based models. For example, Van Segbroeck et al. [23] observe a correlation between agents' behaviors and their degrees

Table 3. Widths of uncertain regions and *UR*. In each cell, the upper entry is the width of uncertain regions, the lower entry is *UR*, and values within parenthesis are standard deviations.

Network type	100 agents	200 agents	500 agents	800 agents	1000 agents
Small-world	0.2716 (0.0352)	0.3141 (0.0394)	0.3673 (0.0411)	0.3752 (0.0332)	0.387 (0.0297)
	0.9495 (0.0088)	0.95 (0.0075)	0.943 (0.0057)	0.948 (0.0146)	0.948 (0.0041)
Scale-free	0.2455 (0.0307)	0.2535 (0.0556)	0.2932 (0.0375)	0.2988 (0.0448)	0.3045 (0.0532)
	0.9555 (0.0076)	0.9595 (0.0051)	0.956 (0.005)	0.956 (0.0069)	0.96 (0.0086)
Star	0.1012 (0.0228)	0.097 (0.0183)	0.0943 (0.0188)	0.1004 (0.0192)	0.0984 (0.0219)
	0.981 (0.0064)	0.986 (0.005)	0.9855 (0.0049)	0.985 (0.0051)	0.9855 (0.005)

in the network. Considering the correlation between agents' neighbors and average shortest path length, Wagner [25] finds that coordinating signals can easily emerge in small-world networks. And Santos et al. [18] discover that cooperation depends on the intricate ties between individuals in scale-free networks.

Apart from these findings, our work, unifying different types of networks, shows that the average shortest path length has a negative correlation with the speed of linguistic categorization. Since average degree, shortcuts, and level of centrality all contribute to the average shortest path length, all of them can collectively affect linguistic categorization. In addition, the clustering coefficient seems not to be crucial as claimed in other studies (e.g., [23]); compared with other networks having high clustering coefficients, the star network, having a low clustering coefficient, can also efficiently affect linguistic categorization. Furthermore, based on the category game, we find that the number of shared words (linguistic categories) is affected by the level of centrality of the network. Further analysis is needed to better evaluate the correlation of level of centrality, population size, and number of shared linguistic categories.

Apart from social factors, nonuniform *dmin* in this language game can also influence linguistic categorization. As discussed in [4], nonuniform *dmin* helps explain the universal color categorization patterns across languages. In addition, as observed in [7], under the same social constraints, the category game behaves differently from the naming game in forming social clusters. Without clarifying linguistic components, the game-theory-based models fail to reveal these in-depth correlations. This point echoes the necessity of involving various linguistic features to analyze the roles of complex networks in language evolution. The general framework advocated in this article—viz., extending language games by incorporating linguistic components other than lexical items, and by studying the dynamics of such extended games in a fully connected network and then in networks with different social constraints—will become a productive procedure to comprehensively evaluate the effect of sociocultural factors on language evolution.

Acknowledgments

This work is partly supported by the EU under RD contract IST-1940 (ECAgents). Gong acknowledges support from the Alexander von Humboldt Foundation in Germany and the Society of the Scholars in the Humanities at the University of Hong Kong.

References

1. Atshogs, Y. V. (2003). *Research on mixing of Tibetan and Chinese in Daohua and relative languages deep-contact study* (in Chinese). Ph.D. dissertation. Nankai University, Tianjin, China.
2. Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.
3. Baronchelli, A., Felici, M., Loreto, V., Caglioti, E., & Steels, L. (2006). Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment*, P06014. Available at http://arxiv.org/abs/physics/0509075v2.
4. Baronchelli, A., Gong, T., Puglisi, A., & Loreto, V. (2010). Modeling the emergence of universal categorization. *Proceedings of the National Academy of Sciences*, *107*(6), 2403–2407.
5. Dall'Asta, L., Baronchelli, A., Barrat, A., & Loreto, V. (2006). Nonequilibrium dynamics of language games on complex networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, *74*(3), 036105. Available at http://arxiv.org/abs/physics/0607054v1.
6. Fisher, M. E., & Barber, M. N. (1972). Scaling theory for finite-size effects in the critical region. *Physical Review Letters*, *28*(23), 1516–1519.
7. Gong, T., Puglisi, A., Loreto, V., & Wang, W. S.-Y. (2008). Conventionalization of linguistic categories under communicative constraints. *Biological Theory: Integrating Development, Evolution, and Cognition*, *3*(2), 154–163.
8. Gong, T. (2009). *Computational simulation in evolutionary linguistics: A study on language emergence* (Frontiers in Linguistics, Monograph IV). Taipei: Institute of Linguistics, Academia Sinica.

9. Kalampokis, A., Kosmidis, K., & Argyrakis, P. (2007). Evolution of vocabulary on scale-free and random networks. *Physica A: Statistical Mechanics and Its Applications*, *379*(2), 665–671.

10. Ke, J.-Y., Gong, T., & Wang, W. S.-Y. (2008). Language change and social networks. *Communications in Computational Physics*, *3*(4), 935–949.

11. Lenaerts, T., Jansen, B., Tuyls, K., & De Vylder, B. (2005). The evolutionary language game: An orthogonal approach. *Journal of Theoretical Biology*, *235*(4), 566–582.

12. Loreto, V., & Steels, L. (2007). Social dynamics: Emergence of language. *Nature Physics*, *3*, 758–760.

13. Lupyan, G., & Rick, D. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*(1), e8559.

14. Mézard, M., Parisi, G., & Virasoro, M. (1987). *Spin glass theory and beyond*. New York: World Scientific.

15. Mukherjee, A., Tria, F., Baronchelli, A., Puglisi, A., & Loreto, V. (2011). Aging in language dynamics. *PLoS ONE*, *6*(2), e16677.

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

17. Puglisi, A., Baronchelli, A., & Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences*, *105*(23), 7936–7940.

18. Santos, F. C., Pacheco, J. M., & Lenaerts, T. (2008). Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proceedings of the National Academy of Sciences*, *103*(9), 3490–3494.

19. Skyrms, B. (2009). Evolution of signaling systems with multiple senders and receivers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1518), 771–779.

20. Steels, L. (2000). Language as a complex adaptive system. In *Parallel problem solving from nature: PPSN-VI* (pp. 17–26). Berlin: Springer-Verlag.

21. Steels, L. (2001). Grounding symbols through evolutionary language games. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 211–226). Berlin: Springer-Verlag.

22. Steels, L. (2004). Constructivist development of grounded construction grammars. In D. Scott, W. Daelemans, & M. Walker (Eds.), *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 9–16). Barcelona, Spain: Association for Computational Linguistics.

23. Van Segbroeck, S., de Jong, S., Nowé, A., Santos, F. C., & Lenaerts, T. (2010). Learning to coordinate in complex networks. *Adaptive Behavior*, *18*(5), 416–427.

24. Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, *167*(1–2), 206–242.

25. Wagner, E. (2009). Communication and structured correlation. *Erkenntnis*, *71*, 377–393.

26. Wang, W. S.-Y. (2003). Language is a complex adaptive system. *Journal of Tsinghua University (Philosophy and Social Science)*, *21*(6), 5–13.

27. Watts, D. J. (1999). *Small worlds: The dynamics of networks between order and randomness*. Princeton, NJ: Princeton University Press.