

THE EFFECT OF SOCIAL POPULARITIES ON LINGUISTIC CATEGORIZATION

TAO GONG

*Department of Linguistics, Max Planck Institute for Evolutionary Anthropology,
Deutscher Platz 6, Leipzig, 04103, Germany*

This paper adopts the category game model that simulates the coevolution of categories and their word labels to explore the effect of social structure on linguistic categorization. Instead of detailed social connections, we adopt social popularities, the probabilities with which individuals participate into language games, to denote quantitatively the general characteristics of social structures. The simulation results show that a certain degree of social scaling could accelerate the categorization process, while a much high degree of social scaling will greatly delay this process.

1. Introduction

Social structures abound in human societies, primate communities, and colonies of some other species (e.g., ants), and social competences (e.g., coalition and competition) are also found in some non-human species. Some of these social factors, existing prior to human language, could have casted their influence on language evolution (Dunbar, 1996). Sociolinguists have already explored the influence of social factors on language change. For instance, Thomason and Kaufman (1988) have illustrated that the social facts of particular contact situations mainly determine the contact-induced language change, and Labov (2001) has shown that language-external factors such as social networks, identity, and gender could determine linguistic variations. In addition, modeling social systems as complex networks can offer new insights on social science research, and this approach has been recently adopted to study historical linguistic change (e.g., Bhattacharjee, 2003) and other sociological phenomena (e.g., Malsch & Schulz-Schaeffer, 2007). In this line of research, the earlier studies adopted simple networks to represent human communities, such as row (Livingstone, 2001), lattice (e.g., Nettle, 1999), and ring (e.g., Kirby, 2000). And later on, some complex networks revealed from sociological studies, such as small-world and scale-free networks, were considered (e.g., Ke, 2004), in which an individual is a node and a communication among individuals is an

edge connecting nodes. Many simulations have shown that the longer the social distance among individuals, the weaker the influence these individuals could cast upon each other, and that the more social connections an individual has, the more influential this individual becomes to affect others.

This way of denoting social relations by particular connections among individuals is reasonable in large-scale communities. In small-scale societies, however, since any individual may directly interact with anyone else, it is unrealistic to define actual connections among individuals. One way to overcome this is to adopt a weighted, fully-connected network, but it is still difficult to specify the actual weights of edges based on the frequency information obtained from empirical evidence. Instead of local, particular connections, a global way to denote quantitatively the general characteristics of social structures is necessary. Considering this, in this paper, we define an individual's *social popularity* as the probability with which this individual participates into communications with others, and use the distribution of all individuals' social popularities to reflect the characteristics of the whole structure. Based on the category game model (Puglisi et al., 2008) that simulates the categorization process involving the evolution of both lexical items and semantic categories, we conduct a simulation study under different distributions of social popularities to explore the effect of social factors on linguistic categorization and discuss the role of social structures in language evolution.

The rest of the paper is organized as follows: Section 2 reviews the category game; Section 3 introduces the power-law distributions of social popularities; Section 4 discusses the simulation results; and finally, Section 5 evaluates our way of characterizing social structures.

2. The Category Game

Semantic categories are culture-dependent conventions shared by individuals for understanding the environment (Gardner, 1985). Their emergence may undergo a self-organization process via iterated interactions among individuals (Lakeoff, 1987). The category game model (Puglisi et al., 2008) theoretically simulates the emergence of common categories that share similar semantic boundaries and identical lexical labels.

This model involves a population of N artificial individuals. Starting from scratch, these individuals, through iterated games, dynamically generate a highly shared pattern of linguistic categories to distinguish stimuli from a perceptual channel. For the sake of simplicity and not losing generality, the perceptual channel is represented by the interval $[0, 1)$, and each stimuli a real number in

this continuous space. A categorization pattern is a partition of the interval $[0, 1)$ into sub-intervals, or *perceptual categories*. Individuals have dynamical inventories of form-meaning associations that link categories with their word labels and evolve during iterated games. In each game, two players (a speaker and a listener) are selected from the population and a scene of $M (\geq 2)$ stimuli chosen from the interval $[0, 1)$ is presented, any two of the stimuli cannot appear at a distance smaller than d_{min} , and one stimulus in this scene is the speaker's topic in this game for the listener to perceive. Figure 1(a) illustrates the game procedure based on two games and Figure 1(b) illustrates the categorization process shown in this model.

As shown in Figure 1(b), all individuals initially have only a perceptual category $[0, 1)$ with no associated words. In the first phase of the evolution, the pressure for discrimination makes the number of perceptual categories increase, and many different words are adopted by different individuals for categories having similar boundaries. Such synonymy reaches a peak and then dries out. When on average only one word is recognized by the whole population for each perceptual category, the second phase intervenes, during which words expand their dominion across adjacent perceptual categories, joining these categories into *linguistic categories*. The coarsening of these categories becomes slower and slower, with a dynamical arrest analogous to the physical process in which super-cooled liquids close to the glass transition (Mézard et al., 1987). In this long-lived, almost stable phase, usually after 10^4 games per individual, the categorization pattern has a degree of sharing between 90% and 100% and remains stable for 10^5 to 10^6 games per individual. If one waits for a much longer time, the number of linguistic categories is observed to drop down, caused by the slow diffusion of category boundaries that ultimately takes place due to small size effects. In this study, we focus on the shared pattern in the stable phase between 10^4 and 10^6 games per individual.

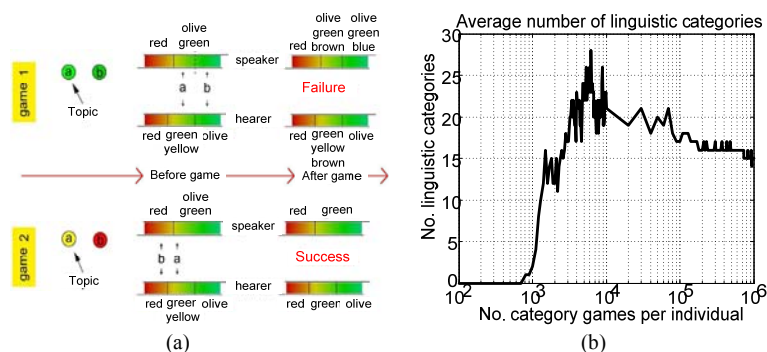


Figure 1. (a) Two examples of the category game (adapted from Puglisi et al., 2008). The round objects are stimuli presented, among which the topics are pointed. The colorful banners represent individuals' perceptual channels, separated by bars into perceptual categories, whose inventories of words are listed above or below. In game 1, since the two stimuli fall into the same category, the speaker discriminates the topic ("a") by creating a new boundary in his/her rightmost perceptual category at the position $(a+b)/2$. Two new categories are created, both inheriting the word-inventory ("green" and "olive") of their parent category, and two new words are invented respectively for these new categories ("brown" and "blue"). Then, the speaker browses the list of words associated to the category that contains the topic. If a previous successful game using this category occurred, its winning word is sent to the listener; otherwise, the newly created word for this category ("brown") is sent. Since the listener does not have this word in his/her inventory, he/she cannot understand it, and the game *fails*. Then, the speaker points at the topic for the listener to discriminate, and the listener adds the speaker's word to the inventory of his/her corresponding category. In game 2, the topic "a" is already discriminated by the speaker in a perceptual category whose winning word is "green". Then, "green" is sent to the listener, who also knows this word and points at the topic contained in his/her corresponding category. This game is *successful*. Then, both individuals eliminate all competing words in their used perceptual categories whose boundaries might not match exactly, but leave the word "green" only. If ambiguity arises (the speaker's word is associated to more than one category that contains the topic), the listener takes a random choice. (b) The categorization process in a random category game simulation, $N=50$ and $d_{min}=0.01$, indicated by the average number of linguistic categories per individual over 50 runs under the same settings.

3. The Representative Distribution of Social Popularities

Sociological research has discovered that instead of uniformity, in many social and linguistic phenomena, such as sexual contact (Lijeros et al., 2003), rumor spread (Moreno et al., 2004), and size ranking of language families (Stauffer et al., 2006), the elements and interactions among them follow a power-law relation (Newman, 2003, 2005). Such a relation is also characteristics in many self-organizing systems, and factors such as preferential attachment (Barabási & Albert, 1999) can lead to this relation in those systems. A power-law relation of two quantities x and y is defined as $y = ax^{-\lambda}$, where a is a scaling parameter, x represents an element or interaction in a given system, y calculates the frequency of this element or interaction, and λ distinguishes different distributions. As summarized by Newman (2003), the λ values in many real-world power-law distributions lie in the interval $[0.0 \ 3.0]$. For instance, it is 1.0 in the frequency ranking of words, 2.0 in the email exchange network, and 3.0 in the citation network.

In our study, we let social popularities follow power-law distributions: x is individual index (rank) from 1 to N , y calculates the popularity for an individual with index x to participate in communications, and a is set to make sure the sum of all probabilities is 1.0. We choose some sampling λ values: 0.0, 1.0, 1.5, 2.0, 2.5, and 3.0. When λ equals 0.0, all individuals have the same probability to communicate with others. This resembles the random case in which individuals

are communicating randomly with each other. As λ increases, individuals having smaller indices become more popular than those having bigger indices, and they tend to communicate much frequently with each other but occasionally with those having bigger indices. Notice that there is a mathematical relation between the λ in the power-law, rank-frequency distribution used in this model and λ' in the power-law, cumulative distributions of node degrees: $\lambda' = 1 + 1/\lambda$.

4. The Simulation Results

In our simulation, $N=50$, $d_{min}=0.01$, individuals conduct 10^6 category games per individual, and their popularities follow the power-law distributions with the sampling λ values. We define two indices to evaluate the categorization process: a) *Overlap*, calculating the average degree of alignment of linguistic categories among individuals, a high value of which indicates that individuals tend to develop categories having similar boundaries; b) *Understanding Rate (UR)*, calculating the percentage of successful category games in all pairs of individuals, a high value of which indicates that individuals tend to use identical labels to describe stimuli for categories with similar boundaries.

Figure 2 shows the simulation results. High *Overlap* of linguistic categories and *UR* are obtained after 10^6 category games per individual. Compared with the random case ($\lambda=0.0$, the solid line), social popularities cast an influence on the categorization process. With the increase in λ , under a fixed number of category games, *Overlap* of perceptual categories becomes smaller. When λ is bigger than 1.5, *Overlap* of linguistic categories starts to drop greatly. Meanwhile, when λ increases from 0.0 to 1.5, the increase in *UR* occurs earlier than that in the random case, which suggests that a certain degree of social scaling can accelerate the emergence of a common set of categories with similar boundaries and identical word labels; when λ exceeds 1.5, however, the increase in *UR* occurs much later, which indicates that a high degree of social scaling could actually delay the emergence of a common set of categories. The simulation results are similar when $N=100$ or $N=200$, and as shown in (Puglisi et al., 2008), a decrease in d_{min} cannot greatly change the categorization process.

These results illustrate a boundary λ value (around 1.0) in power-law distributed social popularities, below which categorization is accelerated, but beyond which this process is delayed. An increase in λ makes individuals with smaller indices have more opportunities to take part in category games, and update and spread their category patterns to others, thus helping to accelerate categorization in the whole population. However, a high λ value will cause individuals with bigger indices to lack opportunities to develop their category

patterns and align their patterns towards those of the popular ones. This delays the increase in *Overlap* and *UR*, and more category games are needed for the whole population to achieve a common set of categories. Due to these two aspects of influence, a power-law distribution with an intermediate λ value becomes a compromise for maintaining both a certain degree of social scaling and a common set of linguistic categories. In addition, considering the relation between rank-frequency and cumulative node degree distributions, $\lambda' = 1 + 1/\lambda$, the boundary λ value (1.0) in the former type of power-law distributions roughly corresponds to the λ' value (2.0) in latter type of power-law distributions. Seen from Newman (2003), the power-law distributions of many language-related social activities have their λ' values around 2.0, such as the email exchange (2.0) and the telephone call (2.1) networks. This provides an empirical support for the boundary λ value shown in our study.

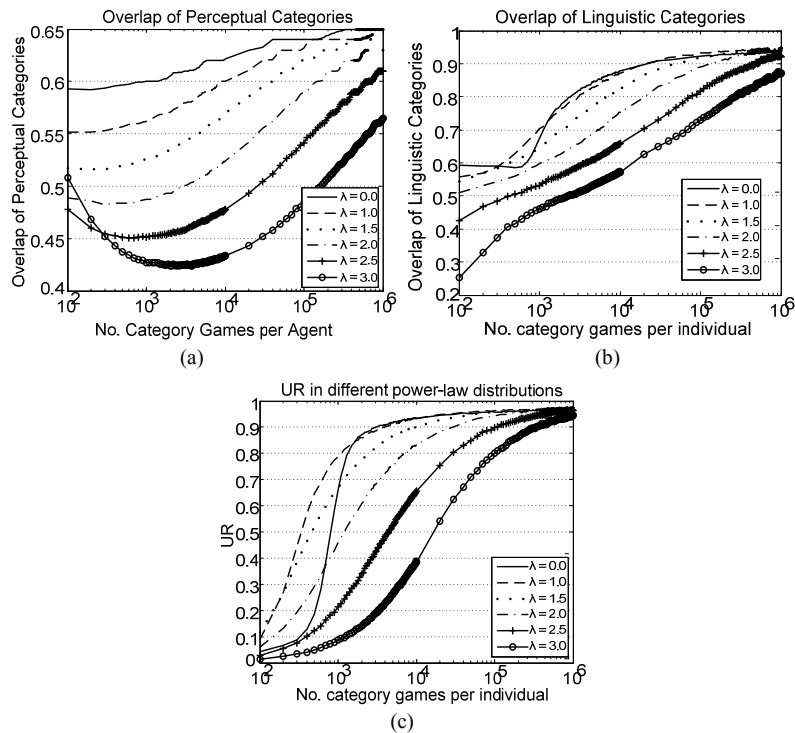


Figure 2. The *Overlap* of perceptual categories (a), *Overlap* of linguistic categories (b), and *UR* (c) in different power-law distributions. Each point is calculated using 20 runs under the same settings.

5. Discussions and Conclusions

Language evolves primarily via social contact among a finite number of individuals. Based on the category game model, we show that a certain degree of social scaling can efficiently develop a common set of semantic categories among individuals, while too much social scaling actually reduces this process. Instead of particular social connections, we adopt the concept of popularity, first used by Nettle (1999), and use the distributions of individual popularities to represent the characteristics of the whole structure. This approach helps summarize the general features shared in various kinds of social structures, requires only a few probability parameters whose values can be obtained from some empirical evidence, and allows statistical analyses to examine the results and provide quantitative understanding on human language, social factors, and their mutual interactions. As shown in Newman (2003), apart from power-law distributions, there are other topological features that are also important in characterizing social communities, and different types of network typology may share functional equivalence on linguistic tasks. The effects of other network typology on language evolution can be studied as well using the same approach. In addition, the category game model is restricted to lexical evolution. To better understand the social structure effect on language evolution, complex language models that concern both lexical and syntactic evolutions should be considered. A preliminary study, based on a lexicon-syntax coevolution model, has shown some similar results about the intermediate λ value in power-law distributions (Gong et al., 2008), which indicates that the power-law relations in a community of individuals is an *independent* factor to influence the average successful rate of language games among individuals.

Acknowledgements

The author acknowledges the support from the Alexander von Humboldt Foundation in Germany, and thanks Prof. Vittorio Loreto and Dr. Andrea Puglisi from Sapienza Universita' di Roma, and Dr. Andrea Baronchelli from Universitat Politecnica de Catalunya for their comments on this work.

References

- Barabási, A-L. & Albert, R. (1999). Emergence of scaling in random networks, *Science*, 286, 509-512.
- Bhattacharjee, Y. (2003). From heofonum to heavens. *Science*, 303, 1326-1328.
- Dunbar, D. (1996). *Grooming, gossip, and the evolution of language*. Cambridge, MA: Harvard University Press.

- Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution*. New York: Basic Books.
- Gong, T., Minett, J. W., & Wang, W. S-Y. (2008). Exploring social structure effect on language evolution based on a computational model. *Connection Science*, 20, 135-153.
- Ke, J-Y. (2004). *Self-organization and language evolution: System, population and individual*. Doctoral Dissertation, City University of Hong Kong.
- Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight (Ed.), *The evolutionary emergence of language: Social function and the origins of linguistic form* (pp. 303-323). Cambridge: Cambridge University Press.
- Labov, W. (2001). *Principles of linguistic change: Social factors*. Oxford, UK: Basil Blackwell.
- Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago, IL: University of Chicago Press.
- Lijeros, F., Edling, C. R., & Amaral, L. A. N. (2003). Sexual networks: Implications for the transmission of sexually transmitted infections. *Microbes and Infection*, 5, 189-196.
- Livingstone, D. (2001). The evolution of dialect diversity. In A. Cangelosi and D. Parisi (Eds.), *Simulating the evolution of language* (pp. 99-117). London: Springer-Verlag.
- Moreno, Y., Nekovee, M., & Pacheco, A. F. (2004). Dynamics of rumor spreading in complex networks. *Physical Review E*, 69, 1-7.
- Stauffer, D., Schulze, C., Lima, F. W. S., Wichmann, S., & Solomon, S. (2006). Non-equilibrium and irreversible simulation of competition among languages. *Physica A*, 371, 719-724.
- Malsch, T. & Schulz-Schaeffer, I. (2007). Socionics: Sociological concepts for social systems of artificial (and human) agents. *Journal of Artificial Societies and Social Simulation*, 10. <http://jasss/soc/surrey.ac.uk/10/1/11.html>.
- Mézard, M., Parisi, G., & Virasoro, M. (1987). *Spin glass theory and beyond*. New York: World Scientific.
- Nettle, D. (1999). Using social impact theory to simulate language change. *Lingua*, 108, 95-117.
- Newman, M. E. J. (2003). The structure and function of complex networks, *SIAM Review*, 45, 167-256.
- Newman, M. E. J. (2005). Power laws, distributions and Zipf's law. *Contemporary Physics*, 46, 323-351.
- Puglisi, A., Baronchelli, A., & Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academic of Sciences*, 105, 7936-7940.
- Thomason, S. G. & Kaufman, T. (1988). *Language contact, creolization, and genetic linguistics*. Los Angeles, CA: University of California Press.