Environmental Impact on Evolving Language Diversity

Gregory Furman, Geoff Nitschke
FRMGRE001@myuct.ac.za,gnitschke@cs.uct.ac.za
Department of Computer Science
University of Cape Town, South Africa

ACM Reference Format:
Gregory Furman, Geoff Nitschke. 2021. Environmental Impact on Evolving Language Diversity. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3459428

EXTENDED ABSTRACT

There has been a significant amount of research on computational modeling of language evolution to understand the origins and evolution of communication [3, 11, 14, 15]. However, there has been relatively little computational modeling of environmental factors influencing the evolution of linguistic diversity and thus the emergence and merging of dialects [1]. Using evolutionary agent-based simulation [3], this study investigates environmental factors influencing the emergence of linguistic diversity in an evolving agent-based language simulation. We used iterative agent-based *naming-game* [15] simulations to evaluate the impact *resources*, *population*, and *environment* size have on evolving language diversity [10]. A specific aim was to investigate thresholds (*tipping-points*) in factors that cause significant changes to linguistic diversity in populations.

Methods and Experiments

Experiments initialized a *Q*x*Q* bounded grid with a random distribution of agent-resource combinations. With uniform randomness, each agent was assigned an arbitrary language-term for five resource types in the environment. Each term was randomly selected to be between 3-9 ASCII characters in length. Resource value was determined by a *Gaussian* random value generator with $\sigma = 2$ and $\mu = 4$, where a resource's value corresponds to fitness received upon an agent consuming *that* resource. Experiments used a *grid world* instantiated with varying numbers of agents, resources and resource types. All agents moved randomly about the grid for their *lifetime* (1000 simulation iterations) during which a variable number of *naming-games* were played. Each agent movement was: North, South, East, or West onto an adjacent unoccupied cell.

A *naming game* was played when an agent moved to a grid-cell adjacent to a resource, and at least one other agent was concurrently adjacent to the resource. All agents then bid a percentage of their fitness (energy) with the highest bidder consuming the resource thus having the payout added to its fitness. All other agents involved in the *naming game* then had their terms for the resource-type being bid over set to that of the winner's term. Thus, the highest bid was deducted from winning agent and paid out equally to all non-winning agents. Bid value was determined by an *Artificial Neural Network* (ANN) controller evolved via NEAT [1] [13].

Each ANN controller had 9 inputs with the first 7 taking in all surrounding agents terms for *that* resource, the 8th receiving the payout of the resource being bid over, and the 9th taking an agent's current fitness. ANN output was a decimal value between 0 and 1 indicating the proportion of an agent's fitness to bid. The fitness function was thus to maximize resources consumed (fitness gained) via bidding specific amounts in naming-games. A population of 500 ANNs was evolved over 100 generations. Each ANN was initialised with 9 input and 1 output node where, at generation 0, all weighted connections were set to 1. NEAT evolved hidden layer connectivity keeping ANN input-output layers fixed. Each generation, all ANNs were evaluated in 50 environments (agent-resource configurations) with average fitness computed over all environments and runs.

Levenshtein Similarity was used to measure linguistic distance between agents with *Hierarchical Complete-Linkage* clustering to classify groups of linguistically similar languages. We used Greenberg's *Linguistic Diversity* (LD) index [6], and *Monolingual Non-weighted Method* for quantifying linguistic diversity. In our experiments, LD is the probability that two agents selected from the population at random will not share a language. As such, 0 indicates *all agents* speak same language (all terms for resources are identical), and 1 indicates *no agents* share a language (all terms are distinct).

Results and Discussion

Ordinary Least Squares Regression [9] indicated a statistically positive relationship between average LD and agent population size. Unit increases (50) in population size saw significant increases (t-test [4], $p<0.01$) in average LD (figure 1, left). Similarly, environment area (grid-cells) was a statistically significant predictor of average LD, where unit increases (1000) led to significant ($p<0.01$) decreases in average LD (figure 1, center). Also, unit increases in resources (500) significantly ($p<0.01$) decreased LD (figure 1, right).

Regression analysis indicates an increasing agent population results in increased LD, while increasing environment size and resources results in decreasing LD. This is enabled by *naming-games* occurring when multiple agents are adjacent to any resource. More resources means potentially more naming-games (language terms shared) between agents which increases with population size. However, LD only increases to a point, since as environment size (figure 1, center) and resources (figure 1, right) increase, then an increased number of naming-games will likely result in convergence on specific language terms, thus decreasing LD in the population. Hence, larger environments reduce the chance of multiple agents

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07…$15.00
https://doi.org/10.1145/3449726.3459428

[1] Experiment parameters and code online: https://tinyurl.com/gecco21-evolang

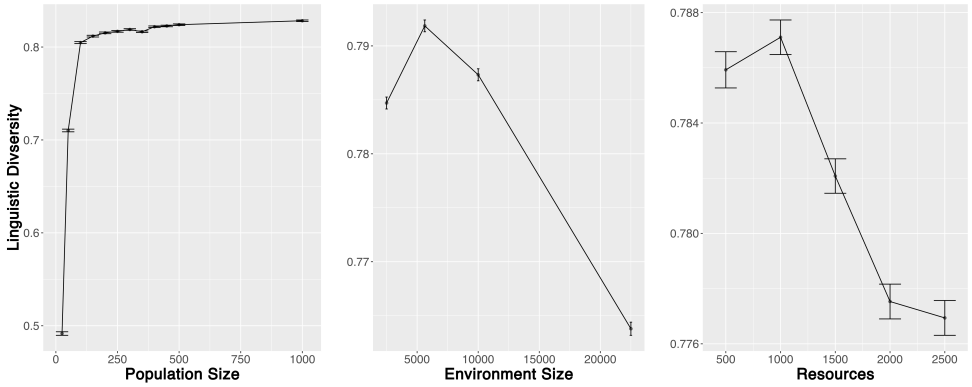

Figure 1: Left: Average *Linguistic Diversity* (LD), averaged over all environments and resource amounts. Center: LD per environment (averaged for all resources and populations). Right: LD per resources (averaged for all environment and populations).

being adjacent to resources and thus interacting (fewer naming-games), which results in a decreasing LD in the population (figure 1, center). Whereas, small environments restrict agent movement (agents and resources cannot occupy the same grid-cell), resulting in fewer naming-games, also resulting in a lower LD in the population.

Notably, LD increases between environment sizes of 2500 to 5625 ($p<0.01$) (figure 1, center), and significantly decreases thereafter ($p<0.01$), indicating an environment size *threshold* where many naming-games facilitate a high LD in the population. A similar phenomenon was observed for resources (figure 1, right), where LD increases between 500 and 1000 resources insignificantly ($p>0.05$) and thereafter decreases significantly ($p<0.01$).

For small environments, agent movement is restricted by *physical barriers* of other agents and resources, thus reducing the number naming-games and limiting LD increase. Though as environment size increases, agent movement increases and more agents concurrently move adjacent to resources thus increasing naming games and LD in the population. Further increased environment size results in greater spread between resources and agents, decreasing the chance that agents concurrently move adjacent to resources, thus decreasing the number of naming-games and LD in the population. As in natural language evolution [2, 7, 12], small environments are analogous to those with many physical barriers, thus inhibiting language sharing but encouraging increased linguistic diversity due to multiple (geographically isolated) dialects in the environment. However, such LD only increases to a point (as a function of population size), given that the number of language (agent) interactions is limited by population size and made less likely by larger environments, thus driving down LD in the population [1].

This is supported by figure 1 (left), indicating that as population size increases, LD in the population increases and then gradually plateaus. Thus, for small population sizes, LD is low due to few naming-games (LD is close to that initialised for the starting population). As population size increases, more naming games are played, thus increasing LD by virtue of agent numbers and diversity of terms. Though such LD increases become negligible for larger populations (>250) ($p>0.05$), since increased naming-games become equated with increased language sharing and thus loss of LD due to convergence on a fixed number of language terms.

In the case of resources, we observe a significant decrease in LD, but for environments containing > 1000 resources. Thus, as with the trend observed for increasing environment size, as resources increase, then LD increases due to increased resource availability enabling more naming games. Though for large resource amounts (>1000) and even greater numbers of naming games, LD once again decreases due to convergence on a fixed number of language terms.

Current research is investigating environment, evolutionary and agent-interaction factors that determine such threshold (tipping-points) observed in these simulations as well as natural language populations [8], where LD increases to a point, and then decreases (for example, changing environment conditions enabling dialects to merge into linguistically less diverse languages [2, 5, 7]).

REFERENCES

[1] T. Arita and Y. Koyama. Evolution of Linguistic Diversity in a Simple Communication System. *Artificial Life*, 4(1):109–124, 1998.
[2] J. Axelsen and S. Manrubia. River Density and Landscape Roughness are Universal Determinants of Linguistic Diversity. *Proceedings of the Royal Society B: Biological Sciences*, 281(1784).
[3] T. Fitch. Empirical Approaches to the Study of Language Evolution. *Psychonomic Bulletin and Review*, 24(1):3–33, 2017.
[4] B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes*. Cambridge University Press, Cambridge, UK, 1986.
[5] M. Gavin and et al. Toward a Mechanistic Understanding of Linguistic Diversity. *BioScience*, 63(7):524–535, 2013.
[6] J. Greenberg. The Measurement of Linguistic Diversity. *Language*, 32(1):109–115, 1956.
[7] T. Honkola and et al. Evolution within a Language: Environmental Differences Contribute to Divergence of Dialect Groups. *BMC Evolutionary Biology*, 18(1):1–15, 2018.
[8] X. Hua, S. Greenhill, M. Cardillo, H. Schneemann, and L. Bromham. The Ecological Drivers of Variation in Global Language Diversity. *Nature Communications*, 10(2047), 2019.
[9] G. Hutcheson. Ordinary Least-Squares Regression. *SAGE Dictionary of Quantitative Management Research*, pages 224–228, 2011.
[10] D. Livingstone. The Evolution of Dialect Diversity. In *Simulating the evolution of language*, pages 99–117. Springer, 2002.
[11] M. Nowak, N. Komarova, and P. Niyogi. Computational and evolutionary aspects of language. *Nature*, 417(1):611–617, 2002.
[12] G. Sankoff. 25 Linguistic Outcomes of Language Contact. *The Handbook of Language Variation and Change*, page 638, 2002.
[13] K. Stanley and R. Miikkulainen. Evolving Neural Networks through Augmenting Topologies. *Evolutionary Computation.*, 10(2):99–127, 2002.
[14] L. Steels. Modeling the Cultural Evolution of Language. *Physics of Life Reviews*, 8(1):339–356, 2011.
[15] L. Steels and A. McIntyre. Spatially Distributed Naming Games. *Advances in Complex Systems*, 1(4):301–323, 1999.