A decentralized route to the origins of scaling in human language

Felipe Urbina\textsuperscript{1} and Javier Vera\textsuperscript{2}

\textsuperscript{1} Centro de Investigación DAI\textsuperscript{T}A Lab, Facultad de Ciencias, Universidad Mayor, Santiago, Chile
\textsuperscript{2} Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
E-mail: furbinaparada@gmail.com and javier.vera@pucv.cl

Received 15 May 2019
Accepted for publication 15 July 2019
Published 6 September 2019

Online at stacks.iop.org/JSTAT/2019/093401
https://doi.org/10.1088/1742-5468/ab38be

\textbf{Abstract.} Zipf’s law establishes that if the words of a (large) text are ordered by decreasing frequency, the frequency versus the rank decreases as a power law with exponent close to $-1$. Previous work has stressed that this statistical pattern arises from a conflict of interests between the participants of communication. The challenge here is to define a computational multi-agent language game, mainly based on a parameter that measures the relative participant’s interests. Numerical simulations suggest that at critical values of the parameter a human-like vocabulary, exhibiting scaling properties, seems to appear. The appearance of an intermediate distribution of frequencies at some critical values of the parameter suggests that on a population of artificial agents the emergence of Zipfian properties partly arises from a self-organized process defined by local interactions between agents.

\textbf{Keywords:} agent-based models, cellular automata, numerical simulations, phase diagrams
1. Introduction

Human vocabularies are constrained by a fundamental statistical principle of organization, the so-called Zipf’s law, which establishes that the frequency of a word decays inversely as a power law of its rank \([1-6]\). More precisely, if the words are ordered by decreasing frequency, the frequency of the \(k\)th word, \(P(k)\), follows

\[ P(k) \sim k^{-\alpha}. \]  

(1)

Remarkably, two important approaches have attempted to describe the origins of scaling in human language. Firstly, within an empirical point of view several works have described the Zipf’s law in language (see, for example \([1-4]\)). Notably, two recent intriguing works have stressed the appearance of scaling regimes in large datasets of written language Perc \([7]\) noticed, on the one hand, that Zipfian properties in a dataset can be understood as an indication of the appearance of large-scale self-organization patterns. In this sense, this paper provided evidence in order to stress the fact that preferential attachment processes, like the Matthew effect \([8, 9]\), played a central role in determining the emergence of scaling in text corpora. On the other hand, using an extremely large corpus of written language \([10]\) found that the word frequency distribution is characterized by two different regimes, and thus extended the seminal observation of \([11]\).

Secondly, in more conceptual terms this scaling law in the distribution of word frequencies has been understood in terms of a dichotomy between high-frequency words, that require little memory effort (like the word ‘the’), and low frequency words, that minimize the disambiguation effort of highly specialized concepts (like the word...
‘computer’). Zipf referred to the lexical trade-off between two competing pressures, *ambiguity* and *memory*, as the *least effort principle*. Within an Information Theoretic analysis, a family of recent works has profound on the understanding of the Zipfian interpretation of scaling properties of language [12–16]. Remarkably, a crucial initial study of Ferrer-i-Cancho and Solé in 2003 [12] showed that Zipf’s law is the outcome of the arrangement of word-meaning associations satisfying the simultaneous interests of both speakers and listeners. Their model results strongly suggested the appearance of a phase transition at a critical least effort stage for both competing pressures. Their study, however, only focused on the emergence of scaling as a simple optimization process operating on a single matrix of word-meaning associations, without any consideration about population structure or communicative interactions between artificial agents.

Here, our work is guided thus by the following question (based on a related proposal of [17]): can artificial populations of agents develop, without any central control, a shared vocabulary satisfying Zipfian properties? The main aim is therefore to define a multi-agent language game [18], according to different levels of word *ambiguity* and *memory* usage. Moreover, the hypothesis is that at some intermediate level of participant’s interests, agents will share a word-meaning mapping exhibiting Zipfian properties. The focus of this paper relies on an abstract decentralized solution in which agents collectively reach shared communication systems without any central control influencing the formation of a human-like language, and only from local conversations involving few participants [19–22].

The model proposed here is based on a prototypical agent-based model for computational studies of language formation, the *naming game* [18, 19, 21, 22], which considers a finite population of agents, each one endowed with a memory to store, in principle, an unlimited number of words. At each discrete time step, a pair of agents, one speaker and one hearer, negotiate words as hypotheses for naming one object. Under the typical dynamics of the *naming game*, the population will share after a finite amount of time a unique word for referring to the object.

The remaining of the article details the dynamics of the agent-based language model. We organize this discussion in three sections. The next section describes the main concepts of the model and the communicative interaction rules. Section 3 presents the simulation protocol for computational experiments. Section 4 describes and illustrates the main results. Section 5 summarizes our work and restates the key challenges of our approach to the formation and evolution of language.

2. Model

2.1. Main elements

The game is played by a finite population of agents $P = \{1, \ldots, p\}$, sharing both a set of words $W = \{1, \ldots, n\}$ and a set of meanings $M = \{1, \ldots, m\}$. A word-meaning mapping can be expressed by a $n \times m$ lexical matrix $L = (l_{ij})$, where $l_{ij} = 1$ if the $i$th word conveys the $j$th meaning, and $l_{ij} = 0$, otherwise. In a related framework, lexical matrices are understood in terms of *language networks* [17], in which there are two disjoint sets of nodes, $W$ and $M$, joined by *edges* establishing word-meaning relationships.
The agent $k \in P$ is endowed in turn with its own word-meaning mapping, expressed by the $n \times m$ lexical matrix $L^k = (l^k_{ij})$, which is known by the agent $k$, and unknown by any other agent $k' \in P \setminus \{k\}$.

Next, two technical concepts are introduced for the agent $k \in P$. The agent $k$ knows the word $i \in W$ if $\sum_{j=1}^{m} l^k_{ij} \geq 1$, that is, the sum of the $i$-th row over the columns of the lexical matrix $L^k$ is greater or equal than 1. The quantity $a^k(i) = \sum_{j=1}^{m} l^k_{ij}$ is called the ambiguity of the word $i$ for the agent $k$.

Given a word $i \in W$, the value of its ambiguity $a^k(i)$ is closely related to the pressures competing in the Zipfian lexical trade-off. In the case of $a^k(i) = 1$, indeed, the speaker’s memory is maximized, since the word $i$ is available only for one meaning $j^* \in M$, whereas the hearer’s disambiguation effort is minimized, since it faces the least effort of disambiguate the word-meaning association. At the opposite case, $a^k(i) = m$, the speaker’s memory is minimized, based on the fact that the word $i$ is associated to every meaning $j \in M$. The hearer’s disambiguation effort, on the contrary, is maximized, since the effort to establish a word-meaning association is maximized.

Some repair strategies for agent’s interactions need the introduction of the following notation: the $j$-th column of the lexical matrix $L^k$ is denoted $L^k_j$. With this notation, the Hadamard product between $L^k_j$ and the canonical vector $e_i$ of dimension $n$ is defined as the entrywise product $L^k_j \odot e_i$. For example, the product between the vectors of dimension $n = 5$, $L^k_{i,j} = (0, 1, 0, 1, 1)^T$ and $e_2 = (0, 1, 0, 0, 0)^T$, is $L^k_{i,j} \odot e_i = (0, 1, 0, 0, 0)^T$.

2.2. Basic interaction rule

The dynamics of the language game is defined by pairwise speaker-hearer interactions at each discrete time step $t \geq 0$. At $t = 0$, each agent $k \in P$ is endowed with a lexical matrix $L^k = (l^k_{ij})$, in which each entry $l^k_{ij}$ is equal to 1 or 0 with probability 0.5. The basic interaction rule is defined by three successive phases at each time step $t$,

- **(phase 1)** a pair of agents is selected uniformly at random: one plays the role of speaker $s$ and the other plays the role of hearer $h$, where $s, h \in P$;
- **(phase 2)** the speaker chooses uniformly at random one column (meaning) $j^* \in \{1, \ldots, m\}$ from its lexical matrix $L^s$. Next, the speaker inspects its own lexical matrix $L^s$ in order to select a word associated to $j^*$, denoted $i^*$. There are two possibilities for this inspection:
  - **(i)** if $\sum_{i=1}^{n} l^s_{ij^*} = 0$, the speaker chooses the word $i^*$ uniformly at random from $W$, and sets $l^s_{ij^*} = 1$ in its lexical matrix $L^s$;
  - **(ii)** otherwise, the speaker calculates the word $i^*$ based on its own lexical interests, that is, based on the of conflict between ambiguity and memory amount.

The speaker transmits the word $i^*$ to the hearer.

- **(phase 3)** The hearer behaves as in the naming game. On the one hand, mutual agreement between the speaker and the hearer involves alignment strategies [22]. On the other hand, if the hearer does not know the word $i^*$, that

https://doi.org/10.1088/1742-5468/ab38be
is, $l^h_{i^*,j^*} = 0$, a repair strategy is established in order to increase the chance of future agreements. More precisely,

(i) if $l^h_{i^*,j^*} \neq 0$, both speaker and hearer update the $j^*$th column of their lexical matrices, by the Hadamard products ($\odot$).

\[
\begin{align*}
L^s_{i^*,j^*} &\leftarrow L^s_{i^*,j^*} \odot e_{i^*}, \\
L^h_{i^*,j^*} &\leftarrow L^h_{i^*,j^*} \odot e_{i^*},
\end{align*}
\]

(ii) otherwise, the hearer establishes a simple repair strategy: it adds 1 to the entry $(i^*j^*)$ of its lexical matrix $L^h$.

### 2.3. Relative interests of speakers and hearers

What would be the minimal adaptation of the basic interaction rule that enables to include at the same time the interests of both speakers and hearers? In order to define relative interests, one feasible solution involves that speakers would prefer to transmit words associated to a relative ambiguity, defined by a simple relationship between its two extreme values of ambiguity: $a^k(i) = 0$ or 1. The solution consists in the following version of the phase 2:

(phase 1) (i) if $\sum_{i=1}^{n} l^s_{ij^*} = 0$, the speaker chooses thus the word $i^*$ uniformly at random from $W$, and sets $l^s_{ij^*} = 1$ in its lexical matrix $L^s$;

(ii) otherwise, the speaker calculates $i^*$ according to the ambiguity parameter $\varphi \in [0, 1]$. Let $random \in [0, 1]$ be a random number. Then,

- if $random \geq \varphi$, the speaker calculates $i^*$ as the least ambiguous word

$$i^* = \min_{\{w: l^s_{wj^*} \neq 0\}} \sum_{i=1}^{n} l^s_{wj^*};$$

- otherwise, the speaker calculates $i^*$ as the most ambiguous word

$$i^* = \max_{\{w: l^s_{wj^*} \neq 0\}} \sum_{i=1}^{n} l^s_{wj^*}.$$ 

#### 2.3.1. Examples

At some time step, let us consider the scenario in which the topic of the interaction is the meaning (column) $j^* = 2$; and the speaker $k \in P$ is endowed with the lexical matrix
A decentralized route to the origins of scaling in human language

\[ L^k = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \]

Therefore, the speaker calculates one of the following words:

- the most ambiguous word (row) \( i^* \) as
  \[ i^* = \arg\max_{\{i \in \{2,3\}} \sum_{j=1}^m l_{ij} = \arg\max_{\{i \in \{2,3\}} \sum_{j=1}^m l_{ij} = 3; \ or \]

- the least ambiguous word (row) \( i^* \) as

**Figure 1.** Size of the effective vocabulary \( \langle V \rangle \) versus ambiguity parameter \( \wp \). On a population of \( p = 128 \) agents, each one endowed with a \( 64 \times 64 \) lexical matrix, it is described the behavior of \( \langle V \rangle \) versus \( \wp \). Two kinds of population organizations are considered. First, a homogeneous mixing of agents, the mean-field aproximation, is studied (denoted \( r = \text{Random} \)). Second, agents are located on a periodic ring of size \( p \). In this case, for each speaker-hearer interaction if the speaker is located at the position \( k \) the hearer is selected uniformly at random from the neighborhood \( (k-r, k+r) \), where the radius \( r \) varies in \( \{1, 2, 4, 8, 16, 32, 64\} \). For all our parameters, the focus here relies on the values after \( 2p \times 10^4 \) speaker-hearer interactions, which average over 10 initial conditions and the last \( 2 \times 10^3 \) steps. One initial condition supposes that each lexical matrix entry is 0 or 1 with probability 0.5. For these measures, the parameter \( \wp \) is varied from 0 to 1 with an increment of 3%. As shown in the figure, the dynamics under the mean-field approximation exhibits a sudden transition that suggest the appearance of scaling properties.
Figure 2. Energy-like function $E_{KL}$ versus $\varphi$. On a population of $p = 128$ agents, each one endowed with a $64 \times 64$ lexical matrix, it is described the behavior of $E_{KL}$ versus $\varphi$ for the dynamics under a homogeneous mixing of agents, the mean-field approximation. The focus relies on the values after $2p \times 10^4$ speaker-hearer interactions, which average over 10 initial conditions and the last $2 \times 10^3$ steps. One initial condition supposes that each lexical matrix entry is 0 or 1 with probability 0.5. For these measures, the parameter $\varphi$ is varied from 0 to 1 with an increment of 3%. As shown in the figure, at the critical value $\varphi_c \approx 0.5$, the energy-like function $E_{KL}$ is minimized.

Figure 3. Appearance of an intermediate frequency distribution in vocabularies. On a population of $p = 128$ agents, each one endowed with a $64 \times 64$ lexical matrix, is shown the behavior of $P(k)$ versus $k$ for the mean-field approximation. For $\varphi = 0.3, 0.8$ and the parameter associated to the power law parameter closest to 1 ($\varphi_c = 0.52$), the figure exhibits the distribution of the number of meanings associated to the $k$-ranked word of the effective vocabulary, $P(k)$, versus $k$ (log - log plot). Black depicted lines indicate least squares fit. The calculations average over ten initial conditions. At the critical parameter $\varphi_c$, the distribution restricted to the words associated at least to one meaning follows $P(k) \sim k^{-\alpha^* = 1.08}$. 

https://doi.org/10.1088/1742-5468/ab38be
A decentralized route to the origins of scaling in human language

\[ i^* = \arg\min_{\{i: l_{ij} \neq 0\}} \sum_{j=1}^{m} t_{ij} = \arg\min_{i \in \{2,3\}} \sum_{j=1}^{m} t_{ij} = 2 \]

and transmits it to the hearer.

3. Measures and simulation protocol

3.1. Two measures

To explicitly describe the dynamics under different lexical interests, two measures are defined: the size of the effective vocabulary [12]

\[ V(t) = \frac{1}{np} \sum_{k=1}^{p} \sum_{i=1}^{m} |i : \sum_{j=1}^{m} t_{ij} > 0| \]  

(2)

where \( \sum_{j=1}^{m} t_{ij} > 0 \) means that the \( i \)th word of the lexical matrix \( L^k \) is being occupied; and the energy-like function \( E_{KL} \), defined as

\[ E_{KL}(\varphi) = d(P(\varphi), P(0)) + d(P(\varphi), P(1)) \]  

(3)

where \( d \) is the symmetric distance defined by the Kullback–Leibler divergence \( KL \) [23]:

\[ d(P(\varphi), P(0)) = KL(P(\varphi), P(0)) + KL(P(0), P(\varphi)) \]

Here, \( P(\varphi) \) denotes the decreasing distribution of frequency meanings for the parameter \( \varphi \). In order to define the probability distribution \( P(\varphi) \), a property is imposed to the ranked frequencies \( p_i^\varphi : \sum_{i=1}^{n} p_i^\varphi = 1 \), for all \( i \in \{1,...,n\} \).

3.2. Parameters of the simulation

The analysis is focused on a homogeneous mixing of agents, the so-called mean-field approximation, in which the population is not structured. In this kind of dynamics, at each time step two agents are selected uniformly at random: one speaker, one hearer. Additionally, as a simple comparison it is described the dynamics on a periodic ring of size \( p = 128 \). For each speaker-hearer interaction, if the speaker is located at the position \( k \) the hearer is selected uniformly at random from the neighborhood \( (k-r, k+r) \), where the radius \( r \) varies in \( \{1, 2, 4, 8, 16, 32, 64\} \). For all our parameters, the focus here relies on the values after \( 2p \times 10^4 \) speaker-hearer interactions, which average over 10 initial conditions and the last \( 2 \times 10^3 \) steps. One initial condition supposes that each lexical matrix entry is 0 or 1 with probability 0.5. For these measures, the parameter \( \varphi \) is varied from 0 to 1 with an increment of 3\%. Each agent is endowed with a \( 64 \times 64 \) lexical matrix.

4. Results

Several aspects are remarkable in the behavior of \( \langle V \rangle \) versus \( \varphi \), as shown in figure 1. In the first place, the dynamics under the mean-field approximation (which is equivalent
to select the radius $r$ at random) exhibits three clear domains. First, $\langle V \rangle$ reaches a value close to 1 for $\varphi < 0.3$. Second, for a critical range of the parameter, $\varphi_c \in (0.3, 0.6)$, it occurs a transition in which $\langle V \rangle$ seems to diminish to 0. Finally, for $\varphi > 0.6$ the dynamics reaches a stationary value $\langle V \rangle \approx 0$.

Qualitatively, the dynamics on one-dimensional rings reproduces the mean-field approximation only for small values of $r$. Indeed, for $r < 16$ the dynamics seems to qualitatively reproduce the three domains described for the mean-field approximation. On the contrary, for $r \geq 16$ the behavior of $\langle V \rangle$ tends to exhibit a linear decay.

One of the most interesting results is summarized by figure 2. At the critical value $\varphi_c \approx 0.5$, the energy-like function $E_{KL}$ is minimized for the dynamics under the mean-field approximation.

5. Discussion

This work summarizes a decentralized route to the origins of scaling properties in a human-like language. The paper describes particularly the influence of a parameter that takes into account agent’s lexical interests during language game dynamics. The appearance of an intermediate distribution of frequencies at some critical values of the parameter suggests that on a population of artificial agents the emergence of scaling partly arises as a self-organized process only from the pressures of local interactions between agents endowed with intermediate levels of lexical interest (for another view on scaling, see figure 3). In some sense, if cooperation is understood as the capacity of selfish agents to forget some of their potential to help one another [24], the emergence of scaling is crucially influenced by the cooperation between agents.

Many extensions of the proposed model should be studied in order to increase the complexity of the language emergence task. A first natural extension should describe more complex ways to define intermediate agent’s interests. A second extension should deal with the miniature artificial language learning paradigm to provide direct experimental evidence for the appearance of Zipfian properties in communicative interactions between real language users optimizing their lexical interests [25].

Further research could involve the problem of why the dynamics over one-dimensional rings involves the maximization of the function $E_{KL}$, unlike the dynamics for the mean-field approximation. One preliminary answer is related to the existente of an optimum range of values for $E_{KL}$.

Acknowledgments

FU thanks FONDECYT Chile for financial support under the Grant 3180227. This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02).
References

[1] Altmann E G and Gerlach M 2016 Statistical Laws in Linguistics (Cham: Springer) pp 7–26
[2] Altmann E G, Dias L and Gerlach M 2017 Generalized entropies and the similarity of texts J. Stat. Mech. 014002
[3] Köhler R, Altmann G and Piotrowski R G (ed) 2005 Quantitative Linguistik Quantitative Linguistics—Ein internationales Handbuch / An International Handbook (Berlin: DeGruyter)
[4] Gerlach M and Altmann E G 2014 Scaling laws and fluctuations in the statistics of word frequencies New J. Phys. 16 113010
[5] Zipf G K 1936 The Psychobiology of Language: an Introduction to Dynamic Philology (London: Routledge)
[6] Zipf G K 1949 Human Behavior and the Principle of Least Effort: an Introduction to Human Ecology (Reading, MA: Addison-Wesley Developers Press)
[7] Perc M 2012 Evolution of the most common english words and phrases over the centuries J. R. Soc. Interface 9 3323–8
[8] Perc M 2014 The Matthew effect in empirical data J. R. Soc. Interface 11 20140378
[9] Perc M 2013 Self-organization of progress across the century of physics Sci. Rep. 3 1720
[10] Petersen A M, Tenenbaum J N, Havlin S, Stanley H E and Perc M 2012 Languages cool as they expand: allometric scaling and the decreasing need for new words Sci. Rep. 2 943
[11] Ferrer-i-Cancho R and Solé R V 2001 Two regimes in the frequency of words and the origins of complex lexicons: Zipf’s law revisited J. Quant. Linguist. 8 165–73
[12] Ferrer-i-Cancho R and Solé R V 2003 Least effort and the origins of scaling in human language Proc. Natl Acad. Sci. USA 100 788–91
[13] Ferrer-i-Cancho R 2005 Zipf’s law from a communicative phase transition Eur. Phys. J B 47 449–57
[14] Ferrer-i-Cancho R 2005 The variation of Zipf’s law in human language Eur. Phys. J. B 44 249–57
[15] Ferrer-i-Cancho R 2006 When language breaks into pieces a conflict between communication through isolated signals and language Biosystems 84 242–53
[16] Ferrer-i-Cancho R and Díaz-Guilera A 2007 The global minima of the communicative energy of natural communication systems J. Stat. Mech. P06009
[17] Solé R V, Cornominas-Murtra B, Valverde S and Steels L 2010 Language networks: their structure, function, and evolution Complexity 15 20–6
[18] Loreto V, Baronchelli A, Mukherjee A, Puglisi A and Tria F 2011 Statistical physics of language dynamics J. Stat. Mech. P04006
[19] Steels L 1995 A self-organizing spatial vocabulary Artif. Life 2 319–32
[20] Steels L 1996 Self-organizing vocabularies Proceedings of Artificial Life V, (Nara, Japan) (Cambridge: MIT Press) pp 179–84
[21] Baronchelli A, Felici M, Caglioti E, Loreto V and Steels L 2006 Sharp transition towards shared vocabularies in multi-agent systems J. Stat. Mech. P06014
[22] Steels L 2011 Modeling the cultural evolution of language Phys. Life Rev. 8 339–56
[23] Bishop C M 2006 Pattern Recognition and Machine Learning (Information Science and Statistics) (Berlin: Springer)
[24] Nowak M 2006 Five rules for the evolution of cooperation Science 314 1560–3
[25] Kanwal J, Smith K, Culbertson J and Kirby S 2017 Zipf’s law of abbreviation and the principle of least effort: language users optimise a miniature lexicon for efficient communication Cognition 165 45–52