Formation of vocabularies in a decentralized graph-based approach to human language

Javier Vera [*]
Pontificia Universidad Católica de Valparaíso, Valparaíso 2340025, Chile

Felipe Urbina [†]
Centro de Investigación DAiTA Lab Facultad de Estudios Interdisciplinarios, Universidad Mayor, Santiago 7560913, Chile

Wenceslao Palma [‡]
Escuela de Ingeniería Informática Pontificia, Universidad Católica de Valparaíso, Valparaíso 2362807, Chile

(Received 19 February 2020; revised 4 January 2021; accepted 26 January 2021; published 15 February 2021)

Zipf's law establishes a scaling behavior for word frequencies in large text corpora. The appearance of Zipfian properties in vocabularies (viewed as an intermediate phase between referentially useless one-word systems and one-to-one word-meaning vocabularies) has been previously explained as an optimization problem for the interests of speakers and hearers. Remarkably, humanlike vocabularies can be viewed also as bipartite graphs. Thus, the aim here is double: within a bipartite-graph approach to human vocabularies, to propose a decentralized language game model for the formation of Zipfian properties. To do this, we define a language game in which a population of artificial agents is involved in idealized linguistic interactions. Numerical simulations show the appearance of a drastic transition from an initially disordered state towards three kinds of vocabularies. Our results open ways to study Zipfian properties in language, reconciling models seeing communication as a global minima of information entropic energies and models focused on self-organization.

DOI: 10.1103/PhysRevE.103.022129

I. INTRODUCTION

This paper arises from two intriguing questions about human language. The first question is: To what extent can language, and also language evolution, be viewed as a graph-theoretical problem? Language is an amazing example of a system of interrelated units at different organization scales. Several recent works have indeed stressed the fact that human language can be viewed as a (complex) network of interacting parts [1–4]. Within the graph-based approach to human language, one may think of word-meaning mappings (that is, *vocabularies*) as bipartite graphs, formed by two disjoint sets: words and meanings [3].

The second question is: What is the nature of the language evolution process that affects the shape of graph-based language representations? To answer this question, we assume that human communication is constrained (at least) by two forces [3]: one that pushes towards communicative success and another that faces the trade-off between speaker and hearer efforts. The first force involves simpler decentralized models of linguistic interactions within populations of artificial agents, endowed with minimal human cognitive features, negotiating pieces of a common language: the so-called *language games* [5–8]. In the simplest language game, the *naming game* [9,10], at a discrete time step, a pair of players (typically one speaker and one hearer) interacts towards agreement on word-meaning associations.

Next, we also consider the communication cost to establish word-meaning mappings. Zipf referred to the lexical trade-off between two competing pressures, *ambiguity* and *memory*, as the *least effort principle* [11,12]: speakers prefer to minimize memory costs, whereas hearers prefer to minimize disambiguation costs. As remarked by several works, an interesting proposal has stated that humanlike vocabularies appear as a phase transition at a critical stage for both competing pressures [13–17]. The appearance of a drastic stage of competing pressures can be understood moreover as an explanation for the empirical Zipf's law, which establishes a dichotomy between low-memory words (such as the word "the") and low-ambiguity words (such as the word "cat"). Within a statistical point of view, text corpora evidence strong scaling properties in word frequencies [18–25]. With this in mind, in our work, the word *Zipfian* must be considered as a sign of an intermediate word-meaning mapping between referentially useless one-word systems and one-to-one word-meaning vocabularies.

The main aim is to address a decentralized approach (based on a previous proposal of two authors of this paper [26]) to the emergence of Zipfian properties in a humanlike language, while players communicate with each other using bipartite word-meaning mappings. To our knowledge, within the *evolutionary linguistics* framework (see, for example, [10]) there is a lack of studies about least-effort communications in decentralized accounts. To structurally characterize changes in the system, our methodology is mainly based on the

[*]javier.vera@pucv.cl
[†]furbinaparada@gmail.com
[‡]wenceslao.palma@pucv.cl

description of vocabularies, arising from both classical statistical mechanics tools and graph-mining techniques. We run numerical simulations over simple population topologies. We apply graph-mining techniques, particularly a clustering notion for bipartite graphs [28].

II. THE MODEL

A. Key concepts on (bipartite) graphs

A *bipartite graph* is a triple $B = (\top, \bot, E)$, where $\top$ and $\bot$ are two mutually disjoint set of nodes, and $E \subseteq \top \times \bot$ is the set of edges of the graph. Here, $\top$ represents the set of *word nodes*, whereas $\bot$ represents the set of *meaning nodes*. We remark that edges only exist between word nodes and meaning nodes. A classical useful tool in graph theory is the matrix representation of graphs. Here, we only consider the *adjacency matrix*. Let us denote by $A = (a)_{uv}$ the adjacency matrix for the (bipartite) graph B. From the bipartite sets $\top$ and $\bot$ representing, respectively, word and meaning nodes, we define the rows of A as word nodes and the columns as meaning nodes, where $(a)_{uv} = 1$ if the word u is joined with the meaning v, and 0 otherwise.

The *neighbors of order 1* of $u \in \top$ are the nodes at distance 1: $N(u) = \{v \in \bot : uv \in E\}$ (if $u \in \bot$, the definition is analogous). Let us denote by $N(N(u))$ the set of nodes at distance 2 from u. The degree $d(u)$ of the node u is simply defined by $d(u) = |N(u)|$. We denote by $d_W^{\max} = \max_{w\in W} d(w)$ the maximum degree for word nodes ($\top$). Analogously, $d_M^{\max} = \max_{m\in M} d(m)$ the maximum degree for meaning nodes ($\bot$).

In social networks, the notion of *clustering* captures the fact that when there is an edge between two nodes (for example, two individuals are friends), they probably have common neighbors. With this, here the *clustering* becomes a simple way to measure correlations between neighborhoods. Based on this point of view, Ref. [28] proposed a clustering coefficient notion for bipartite graphs,

$$cc(u) = \frac{\sum_{v\in N(N(u))} cc(u, v)}{N(N(u))}, \tag{1}$$

where $cc(u, v)$ is a notion of clustering defined for pairs of nodes (in the same set $\top$ or $\bot$),

$$cc(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}. \tag{2}$$

Interestingly, $cc(u, v)$ captures the overlap between the neighborhoods of u and v: if u and v do not share neighbors, $cc(u, v) = 0$, and if they have the same neighborhood, $cc(u, v) = 1$.

To give an overall overview of bipartite clustering for the graph B, the *average bipartite clustering* reads

$$c(B) = \frac{1}{|\top| + |\bot|} \sum_{u\in\top\cup\bot} cc(u). \tag{3}$$

B. Basic elements of the language game

The language game is played by a finite population of participants, $P = \{1, \dots, p\}$, sharing both a set of words, $W = \{1, \dots, \mathcal{N}\}$, and a set of meanings, $M = \{1, \dots, \mathcal{M}\}$. Each player $k \in P$ is endowed with a graph-based word-meaning mapping, $B^k = (\top^k, \bot^k, E^k)$. In our case, B^k is a bipartite graph with two disjoint sets: $\top^k \subseteq W$ (word nodes) and $\bot^k \subseteq M$ (meaning nodes). Each player $k \in P$ only knows its own graph B^k.

Two technical terms are introduced. First, we say that a player $k \in P$ *knows* the word $w \in W$ if $w \in \top^k$. Clearly, this definition is equivalent to the existence of the edge $wm \in E^k$, for some $m \in \bot^k$. Second, the *ambiguity* of the word w, denoted $a(w)$, is defined as its node degree $d(w)$.

C. Language game rules

The dynamics of the language game is based on pairwise speaker-hearer interactions at discrete time steps. At $t \geqslant 0$, a pair of players is selected uniformly at random: one plays the role of *speaker* s and the other plays the role of *hearer* h, where $s, h \in P$. Each speaker-hearer communicative interaction is defined by two successive phases. The first phase involves the selection of a word to transmit a specific meaning. Next, the hearer receives the word-meaning association and both speaker and hearer attempt to *align* their vocabularies. As we can observe, in our model, the first phase introduces a parameter that penalizes having different meanings for a word. This is a bias against ambiguity. The second phase penalizes multiple words for the same meaning and does not penalize having different words for each meaning (a bias against synonymy, not ambiguity).

At each communicative interaction, the speaker s selects one meaning $m^* \in M$. This selection is done uniformly at random from the set M. To transmit m^*, she needs to select some word, denoted w^*. If she does not know a word with the meaning m^*, she selects a word at random from her vocabulary and adds this word-meaning pair to her lexicon B^s.

Then, she *calculates* w^* based on her interests. She behaves according to the *ambiguity parameter* $\wp \in [0, 1]$: with probability 1-$\wp$, w^* is the least ambiguous word,

$$w^* = \min_{w\in\top^s} a(w), \tag{4}$$

while with probability $\wp$, w^* is the most ambiguous word,

$$w^* = \max_{w\in\top^s} a(w). \tag{5}$$

She transmits the word w^* to the hearer.

In turn, the hearer behaves as in the *naming game*. On the one hand, if there is a mutual speaker-hearer agreement (the hearer knows the word w^*), *alignment* strategies appear [10]. On the other hand, a speaker-hearer disagreement (if the hearer does not know the word w^*), involves a *repair* strategy in order to increase the chance of future agreements (that is, for $t' > t$). More precisely, if the hearer knows the word w^*, both speaker and hearer remove all word-meaning pairs wm^* from their vocabularies, where w, respectively, belongs to $\top^s \setminus \{w^*\}$ and $\top^h \setminus \{w^*\}$. By contrary, if the hearer does not know the word w^*, she adds the word-meaning pair w^*m^* to her vocabulary B^h.

III. METHODS

To study the most simple interaction topology (i.e., any two agents picked up randomly can interact), the population

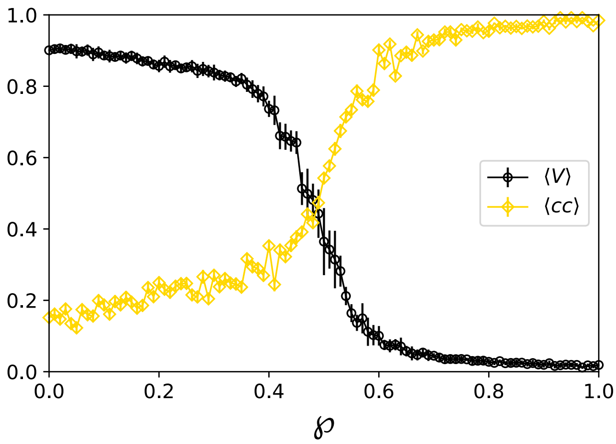

FIG. 1. Measures to describe the formation of vocabularies. Two quantities are shown against the ambiguity parameter $\wp$: average population clustering $cc(t_f)$ and effective vocabulary $V(t_f)$. Averages over 10 realizations. Vertical lines indicate standard deviation. As shown in the figure, three domains for the evolution of bipartite word-meaning mappings tend to appear: full, Zipfian-like, and single-word vocabularies.

of agents is located on the vertices of a complete graph of size $|P| = 100$, typically called the *mean field* approximation. The population shares both a set of $\mathcal{N} = |W| = 128$ words and a set of $\mathcal{M} = |M| = 128$ meanings. Starting from an initial condition in which each player $k \in P$ is associated to a bipartite graph B^k where $B^k_{ij} = 1$ or $B^k_{ij} = 0$ with probability 0.5 (put differently, for each possible edge ij, $i \in W$ and $j \in M$ exist with probability 0.5), the dynamics performs a speaker-hearer interaction at each discrete time step $t \geqslant 0$. More precisely, one speaker $s \in P$ and one hearer $h \in V(s)$ are chosen uniformly at random. The bipartite word-meaning mappings B^s and B^h are then reevaluated according to communicative success. All results consider averages over 10 initial conditions and 3×10^5 time steps (denoted by $\langle\rangle$). We denote by t_f the final time step. The ambiguity parameter $\wp$ is varied from 0 to 1, with an increment of 1%.

IV. RESULTS

A. Formation of three domains for vocabularies

Two key quantities have been analyzed for different values of $\wp$: the *average population clustering cc*,

$$cc = \frac{1}{|P|}\sum_{k\in P} cc(B^k), \tag{6}$$

which captures the average correlation between word neighborhoods, and the (effective) lexicon size at time step t, $V(t)$, defined as [13,26]

$$V(t) = \frac{1}{n|P|}\sum_{k\in P} |\top^k|, \tag{7}$$

where $V(t) = 1$ if $|\top^k| = n$, while $V(t) = 0$ if $|\top^k| = 0$.

Three domains can be noticed in the behavior of $\langle cc\rangle$ versus $\wp$, at t_f, as shown in Fig. 1 (gold diamonds). First, $\langle cc\rangle$ increases smoothly for $\wp < 0.4$, indicating that for this domain, there is a small correlation between word neighborhoods. Full

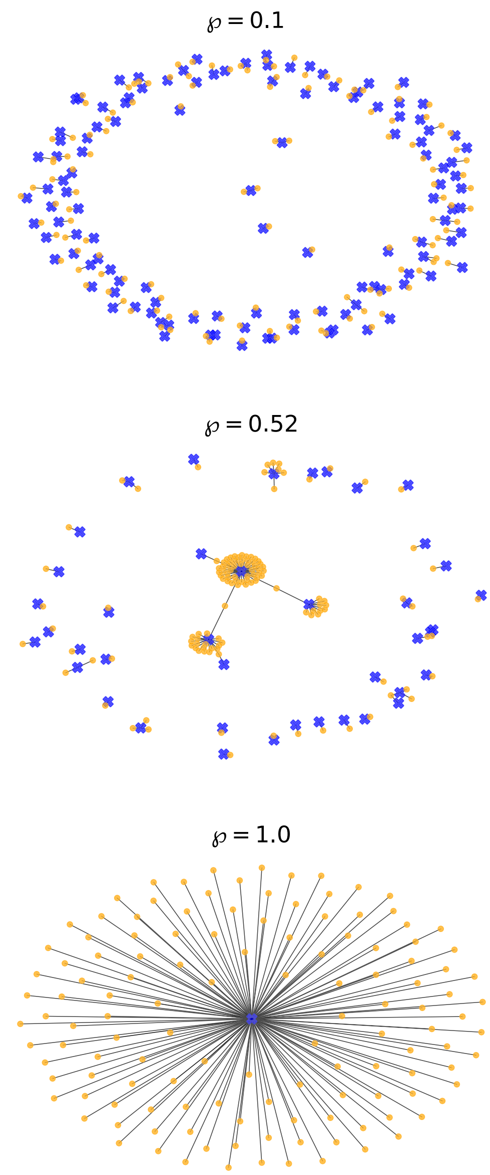

FIG. 2. Phase transition in bipartite graphs. Orange circles indicate meanings, whereas blue x's indicate words. From top to bottom, we choose three bipartite graphs corresponding, respectively, to $\wp$ in $\{0.1, 0.52, 1\}$. Node positions are based on a PYTHON [27] implementation of the Fruchterman-Reingold algorithm [29].

vocabularies are also attained for $\wp < 0.4$. Second, a drastic transition appears at the critical domain $\wp^* \in (0.4, 0.6)$, in which $\langle cc\rangle$ shifts abruptly towards 1. An abrupt change in $V(t_f)$ versus $\wp$ is also found (Fig. 1, black circles) for $\wp^*$. Third, single-word languages dominate for $\wp > 0.6$. The maximum value of $\langle cc\rangle$ indicate that word neighborhoods are completely correlated.

B. Understanding the formation of vocabularies with bipartite graphs

We now shift our focus from graph-based measures towards a holistic level in which we illustrate the formation of vocabularies using bipartite-graph representations. We stress

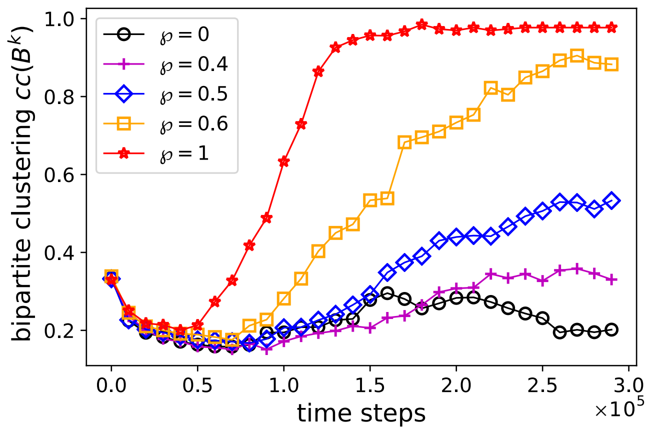

FIG. 3. Average bipartite clustering vs time. For one player $k \in P$, the behavior of average clustering $cc(B^k)$ over time is described. Circles indicate measurements every 10^4 speaker-hearer interactions (starting from $t = 0$). For the initial condition, the player $k \in P$ is associated to a bipartite graph B^k, where $B^k_{ij} = 1$ or $B^k_{ij} = 0$ with probability 0.5. $\wp$ is varied from $\{0, 0.4, 0.5, 0.6, 1\}$.

the fact that our framework based on a language game with players endowed with bipartite word-meaning mappings is able to visualize the structural changes of the three described domains. Figure 2 displays, from top to bottom, the bipartite word-meaning mappings for ambiguity parameters $\wp$ in $\{0.1, 0.52, 1\}$. As expected, there are radical structural changes between bipartite graphs associated to such ambiguity parameters. Full vocabularies are attained for $\wp = 0.1$, located at the hearer-centered phase. Zipfian vocabularies seem to appear for $\wp = 0.52$, where speaker and hearer costs have a similar value. Finally, a single-word vocabulary (that is, one word, several meanings) is exhibited for $\wp = 1$.

C. Language dynamics over time

The *average population clustering* $cc(B^k)$ has been analyzed for different values of $\wp$ over language game dynamics, as shown in Fig. 3, and the *maximum degree* for word nodes $d_W^{\max}$ (black circles) and meaning nodes $d_M^{\max}$ (red stars), as shown in Fig. 4. For this figure, measurements are captured every 10^4 speaker-hearer interactions.

We remark that this dynamics represents only one player $k \in P$. Several aspects are noticeable for the evolution of the average bipartite clustering $cc(B^k)$. In the first place, bipartite clustering curves are superposed until $t < 5 \times 10^4$ time steps, where $cc(B^k)$ seems to reach a local minimum. We notice that for the most speaker-centered situation ($\wp = 1$), bipartite clustering exhibits a drastic transition from 0.2 towards 1 at a critical phase $0.5 \times 10^4 < t < 1.5 \times 10^5$. In turn, for $\wp < 1$, there is a smooth transition starting approximately at $t \sim 10^5$. After that, curves corresponding to different values of the parameter $\wp$ tend to converge to their final values. Particularly, for the critical value $\wp = 0.5$, we can observe an intermediate evolution towards $cc(B^k) \approx 0.5$.

To more deeply understand the graph structural changes over time, we describe the maximum degree for word and meaning nodes at three different parameters $\wp = 0.1, 0.52, 1$. As shown in Fig. 4, the curves corresponding to $d_W^{\max}$ and

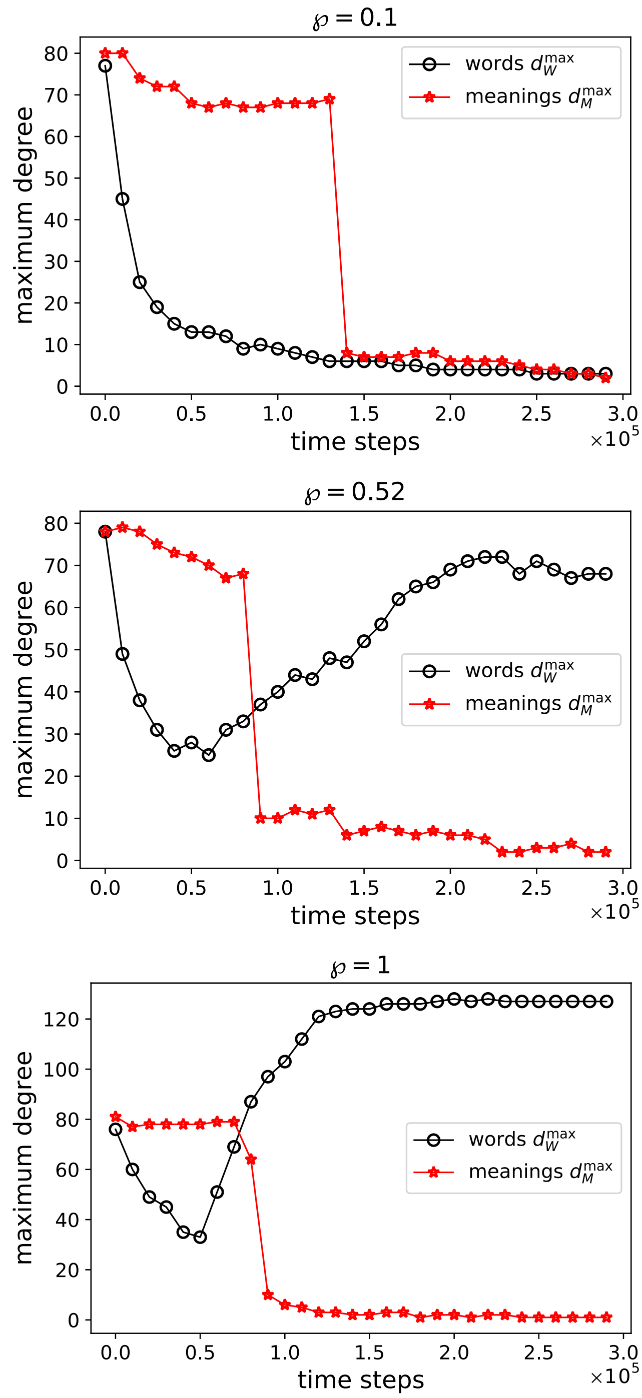

FIG. 4. Maximum degree over time. For one player $k \in P$, the behavior of the *maximum degree for word nodes* $d_W^{\max}$ and the *maximum degree for meaning nodes* $d_M^{\max}$ over time is described. Points indicate measurements every 10^4 speaker-hearer interactions (starting from $t = 0$). For the initial condition, the player $k \in P$ is associated to a bipartite graph B^k, where $B^k_{ij} = 1$ or $B^k_{ij} = 0$ with probability 0.5. $\wp$ is varied from $\{0.1, 0.52, 1\}$.

$d_M^{\max}$ exhibit a drastic transition at $t \sim 10^5$, for $\wp = 1$. On the other hand, for $\wp = 0.1$, only the curve $d_M^{\max}$ abruptly converges to the same value of $d_W^{\max}$. This case indeed supposes a hearer-centered context, in which the player has an almost one-to-one word-meaning mapping. For the humanlike behavior, $\wp = 0.52$, the transition occurring at $t \sim 10^5$ is smoother. This fact suggests that the features of humanlike vocabularies appear early in the dynamics, and not only at the final consensus absorbing state.

V. DISCUSSION

In this paper, we have described a decentralized model of the emergence of Zipfian features in a humanlike language, where agents play language games communicating with bipartite word-meaning mappings. The model suggests the formation of a humanlike vocabulary satisfying Zipfian word-meaning properties. Our central graph-mining tool has been a notion of clustering for bipartite graphs. This function allowed us to suggest that the drastic transition is, in some sense, a qualitative transition in the word's correlations.

To further understand the nature of the described transition, we note a recent proposal [30], reinterpreting an old question about language learning with an alternative approach: if language learning by a child involves setting many parameters, to what extent do all these need to be innate? According to the *principles and parameters* theory [31], children are biologically endowed with a general "grammar" and then the simple exposition to a particular language (for example, *Quechua*) fixes its syntax by equalizing parameters. This debate was illuminated by proposing a statistical mechanics approach in which the distribution of grammar weights (where language is modeled by weighted context-free grammars) evidences a drastic transition. Language learning is, for this proposal, a transition from a random model of grammar parameter weights to the one in which deep structure (that is, syntax) is encountered.

Here, the language learning problem is situated in a decentralized process, with agents negotiating a common word-meaning mapping exhibiting Zipfian scaling properties. Interestingly, our approach can shed light on the debate opened by Ref. [30]. Indeed, our model questioned, first, the fact that language learning is traditionally viewed as an individual process, without any consideration of population structure (in general, *language games* question this fact). Second, we argue that our view pointed out the minimal necessity of cognitive principles for cultural language formation: the least effort principle. We hypothesize that players only need the most basic cognitive features for language learning (and formation) and the rest is an emergent property from the local speaker-hearer interactions. It is interesting to remark that several works have stressed the fact that language formation can be viewed as a *phase transition* within an information-theoretic approach [13–17].

Future work could explore an intriguing hypothesis: Zipfian properties have strong consequences for syntax and symbolic reference. Reference [32] has indeed proposed that Zipf's law is a necessary precondition for full syntax and for going beyond simple word-meaning mappings. They hypothesized, moreover, that the appearance of syntax has been as abrupt as the transition to Zipf's law. This is a goal for future work: to propose a decentralized model in which agents (constrained by specific cognitive features) develop a Zipfian language that acts as a precondition for the abrupt transition to simple forms of syntax (based, for example, on [33]). Another related research line arises from models assuming the interplay between maximization of the information transfer and minimization of the entropy of signals (see, for example, [13,17,34]). As previously remarked, these models evidence a lack of population structure. Current work asks how a community of individuals playing the language game proposed here can minimize the energy functional $\Omega(\lambda)$ of word-meaning mappings. This quantity is formed by the combination of two terms,

$$\Omega_{\wp} = \wp H(R|S) + (1 - \wp) H(S), \tag{8}$$

where $\wp$ is a parameter in [0,1], $H(R|S)$ is the effort for the hearer, and $H(S)$ is the effort for the speaker. In its original form [13], the hypothesis that Zipfian vocabularies should appear at $\wp \approx 0.5$ is assumed, where the efforts of the speaker and the hearer have a similar contribution to $\Omega_{\wp}$.

We may hypothesize, in our language game approach, that reaching global consensus at the critical phase (that is, for $\wp \approx 0.5$) is closely related to the global minima of $\Omega(\lambda)$. This idea opens fascinating ways to study human language, reconciling models seeing communication as a global minima of information entropic energies and models focused on populations self-organizing themselves towards a shared consensus.

ACKNOWLEDGMENTS

F.U. thanks Postdoctoral Grant (ANID-Chile) No. 3180227.

[1] J. Cong and H. Liu, Approaching human language with complex networks, Phys. Life Rev. **11**, 598 (2014).

[2] Y. Gao, W. Liang, Y. Shi, and Q. Huang, Comparison of directed and weighted co-occurrence networks of six languages, Physica A **393**, 579 (2014).

[3] R. V. Solé, B. Corominas-Murtra, S. Valverde, and L. Steels, Language networks: Their structure, function, and evolution, Complexity **15**, 20 (2010).

[4] L. F. Seoane and R. Solé, The morphospace of language networks, Sci. Rep. **8**, 10465 (2018).

[5] V. Loreto, A. Baronchelli, A. Mukherjee, A. Puglisi, and F. Tria, Statistical physics of language dynamics, J. Stat. Mech.: Theory Exp. (2011) P04006.

[6] L. Steels, *The Talking Heads Experiment: Origins of Words and Meanings*, Computational Models of Language Evolution (Language Science Press, Berlin, 2015).

[7] L. Steels, Do languages evolve or merely change? J. Neurolinguist. **43**, 199 (2017).

[8] B. de Boer, Self-organization and language evolution, in *The Oxford Handbook of Language Evolution*, edited by K. R. Gibson and M. Tallerman (Oxford University Press, Oxford, 2011).

[9] A. Baronchelli, M. Felici, V. Loreto, E. Caglioti, and L. Steels, Sharp transition towards shared vocabularies in multi-agent systems, J. Stat. Mech.: Theory Exp. (2006) P06014.

[10] L. Steels, Modeling the cultural evolution of language, Phys. Life Rev. **8**, 339 (2011).
[11] G. K. Zipf, *The Psychobiology of Language: An Introduction to Dynamic Philology* (Routledge, London, 1936).
[12] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley, Cambridge, 1949).
[13] R. Ferrer-i-Cancho and R. V. Solé, Least effort and the origins of scaling in human language, Proc. Natl. Acad. Sci. **100**, 788 (2003).
[14] R. Ferrer-i-Cancho, Zipf's law from a communicative phase transition, Eur. Phys. J. B **47**, 449 (2005).
[15] R. Ferrer-i-Cancho, The variation of Zipf's law in human language, Eur. Phys. J. B **44**, 249 (2005).
[16] R. Ferrer-i-Cancho, When language breaks into pieces a conflict between communication through isolated signals and language, Biosystems **84**, 242 (2006).
[17] R. Ferrer i Cancho and A. Díaz-Guilera, The global minima of the communicative energy of natural communication systems, J. Stat. Mech.: Theory Exp. (2007) P06009.
[18] E. G. Altmann and M. Gerlach, Statistical laws in linguistics, in *Creativity and Universality in Language*, edited by M. Degli Esposti, E. G. Altmann, and F. Pachet (Springer International, Cham, 2016), pp. 7–26.
[19] E. G. Altmann, L. Dias, and M. Gerlach, Generalized entropies and the similarity of texts, J. Stat. Mech.: Theory Exp. (2017) 014002.
[20] *Quantitative Linguistics - An International Handbook*, edited by R. Köhler, G. Altmann, and R. G. Piotrowski (DeGruyter, Berlin, 2005).
[21] M. Gerlach and E. G. Altmann, Scaling laws and fluctuations in the statistics of word frequencies, New J. Phys. **16**, 113010 (2014).
[22] M. Perc, Evolution of the most common English words and phrases over the centuries, J. R. Soc., Interface **9**, 3323 (2012).
[23] M. Perc, Self-organization of progress across the century of physics, Sci. Rep. **3**, 01720 (2013).
[24] M. Perc, The Matthew effect in empirical data, J. R. Soc., Interface **11**, 20140378 (2014).
[25] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc, Languages cool as they expand: Allometric scaling and the decreasing need for new words, Sci. Rep. **2**, 00943 (2012).
[26] F. Urbina and J. Vera, A decentralized route to the origins of scaling in human language, J. Stat. Mech.: Theory Exp. (2019) 093401.
[27] A. A. Hagberg, D. A. Schult, and P. J. Swart, Exploring network structure, dynamics, and function using NetworkX, in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, edited by G. Varoquaux, T. Vaught, and J. Millman, (Pasadena, CA, USA, 2008), pp. 11–15.
[28] M. Latapy, C. Magnien, and N. D. Vecchio, Basic notions for the analysis of large affiliation networks/bipartite graphs, arXiv:cond-mat/0611631.
[29] T. M. J. Fruchterman and E. M. Reingold, Graph drawing by force-directed placement, Software Practice Experience **21**, 1129 (1991).
[30] E. DeGiuli, Random Language Model, Phys. Rev. Lett. **122**, 128301 (2019).
[31] N. Chomsky, *Lectures on Government and Binding: The Pisa Lectures*, Studies in Generative Grammar (DeGruyter, Berlin, 1993).
[32] R. Ferrer-i-Cancho, O. Riordan, and B. Bollobás, The consequences of Zipf's law for syntax and symbolic reference, Proc. Biol Sci. **272**, 561 (2005).
[33] L. Steels, Agent-based models for the emergence and evolution of grammar, Philos. Trans. R. Soc. London Ser. B, Biol. Sci. **371**, 20150447 (2016).
[34] L. Steels and E. Szathmáry, The evolutionary dynamics of language, Biosystems **164**, 128 (2018).