Complex systems approach to the emergence of language

Andrea Baronchelli
Departament de Física i Enginyeria Nuclear,
Universitat Politècnica de Catalunya
Campus Nord, Mòdul B4, 08034 Barcelona, Spain

Ciro Cattuto
Museo Storico della Fisica e Centro Studi e Ricerche Enrico Fermi
Compendio Viminale, 00184 Roma, Italy
and Complex Networks Lagrange Laboratory, ISI Foundation
Viale S. Severo 65, 10133, Torino, Italy

Vittorio Loreto*
Dipartimento di Fisica, Università di Roma "La Sapienza"
p.le Aldo Moro 2, 00185, Roma, Italy
and Complex Networks Lagrange Laboratory, ISI Foundation
Viale S. Severo 65, 10133, Torino, Italy

Andrea Puglisi
Dipartimento di Fisica, Università di Roma "La Sapienza"
p.le Aldo Moro 2, 00185, Roma, Italy

November 9, 2007

Abstract

In this paper we illustrate a statistical physics approach to problem of the emergence of language. In this perspective language is seen as an evolving and self-organizing system, whose components are continuously shaped and (ri)-shaped through the interactions among the individuals of a population. We illustrate several examples corresponding to the early stages of the emergence of a language, namely the emergence of a common lexicon and the emergence of a shared set of linguistics categories. We also discuss an experimental framework related to the World Wide Web where new lexicons and new conventions are emerging trough the collective activity of millions of users. Finally we highlight the potential of such an approach in the interdisciplinary effort to understand the origin of linguistic structures.

*Corresponding author: vittorio.loreto@roma1.infn.it

1 Introduction

One of the first questions the reader could ask while reading this paper is what physics in general, and particularly statistical physics, has to do with the problem of the emergence of language.

Statistical physics has proven to be a very fruitful framework to describe phenomena outside the realm of traditional physics [1]. The last years have witnessed the attempt by physicists to study collective phenomena emerging from the interactions of individuals as elementary units in social structures [2]. In social phenomena the basic constituents are not particles but humans and every individual interacts with a limited number of peers, usually negligible compared to the total number of people in the system. In spite of that, human societies are characterized by stunning global regularities [3]. There are transitions from disorder to order, like the spontaneous formation of a common language/culture or the emergence of consensus about a specific issue. There are examples of scaling and universality. These macroscopic phenomena naturally call for a statistical physics approach to social behavior, i.e., the attempt to understand regularities at large scale as collective effects of the interaction among single individuals, considered as relatively simple entities. This is the paradigm of the Complex Systems science: an assembly of many interacting (and simple) units whose collective (i.e., large scale) behavior is not trivially deducible from the knowledge of the rules that govern their mutual interactions.

Now consider the problem of the emergence of language. It is of course true that if one adopts a static point of view where language is seen as a system frozen at a particular point in time with its sound structure, vocabulary and grammar, there is no place for statistical physics oriented studies. But as linguists begin to get access to more and more data from systematic recordings and the massive volume of text appearing on the World Wide Web, and as they look at new language-like communication systems that have emerged recently - such as text messaging protocols for use with mobile phones or social tagging of resources available on the web - doubts arise whether human communication systems can be captured within a static picture or in a clean formal calculus. The static picture is giving way to a view where language is undergoing constant change as speakers and hearers use all their available resources in creative ways to achieve their communicative goals.

This is the point of view of semiotic dynamics [4] which looks at language as an adaptive evolving system where new words and grammatical constructions may be invented or acquired, new meanings may arise, the relation between language and meaning may shift (e.g. if a word adopts a new meaning), the relation between meanings and the world may shift (e.g. if new perceptually grounded categories are introduced). All these changes happen both at the level of the individual and at the group level, the focus being on the interactions among the individuals as well as on horizontal, i.e., peer to peer, communications. Semiotic dynamics is the sub-field of dynamics that studies the properties of such evolving semiotic systems. In this new perspective, complex systems science turns out to be a natural allied in the quest for the general mechanisms underlying the emergence of a shared set of conventions in a population of individuals.

In semiotic dynamics models assume a population of agents that have only local interactions and carry out some communicative task, such as drawing the attention of another agent to an object in their surroundings by using a name.

Typically agents do not start with a given communication system but must build one up from scratch. The communication evolves through successive conversations, i.e., events that involve a certain number of agents (two, in practical implementations) and meanings. It is worth remarking that here conversations are particular cases of language games, which, as already pointed out in [5, 6], can be used to describe linguistic behavior, even if they can include also non linguistic behavior, such as pointing.

In this paper we explore several modalities of interactions among the individuals of a population corresponding to language games of increasing complexity. We first consider the so-called *Naming Game* (Sect. 2), which possibly represents the simplest example of the complex processes leading progressively to the establishment of human-like languages. It was expressly conceived to explore the role of self-organization in the evolution of language [7, 8] and it has acquired, since then, a paradigmatic role. The original paper [7], focused mainly on the formation of vocabularies, i.e., a set of mappings between words and meanings (for instance physical objects). In this context, each agent develops its own vocabulary in a random private fashion, but agents are forced to align their vocabularies, through successive conversations, in order to obtain the benefit of cooperating through communication. Thus, a globally shared vocabulary emerges as a result of local adjustments of individual word-meaning association.

The next step we consider is the so-called *Category Game* (Sect. 3). In this case one focuses on the process by which a population of individuals manages to categorize a single perceptually continuous channel, each stimulus being represented as a real-valued number ranging in the interval $[0, 1]$. The problem of the emergence of a discrete shared set of categories out of a continuous perceptual channel is a notoriously difficult problem relevant for color categorization, vowels formation, etc. The central result here is the emergence of a hierarchical category structure made of two distinct levels: a basic layer, responsible for fine discrimination of the environment, and a shared linguistic layer that groups together perceptions to guarantee communicative success. Remarkably, the number of linguistic categories turns out to be finite and small, as observed in natural languages.

Finally we conclude by focusing on results coming from an experimental platform: the Web (Sect. 4). Though only a few years old, the growth of the World Wide Web and its effect on the society have been astonishing, spreading from the research in high-energy physics into other scientific disciplines, academe in general, commerce, entertainment, politics and almost anywhere where communication serves a purpose. Innovation has widened the possibilities for communication. Social media like blogs, wikis and social bookmarking tools allow the immediacy of conversation, with unprecedented levels of communication speed and community size. In this perspective the web is acquiring the status of a platform for *social computing*, able to coordinate and exploit the cognitive abilities of the users for a given task. In this sense, it is likely that the new social platforms appearing on the web, could rapidly become a very interesting laboratory for social sciences. Here we report in particular on the temporal growth of tag dictionaries in collaborative tagging systems and discuss the implications for studies on the emergence of language.

2 The Naming Game

In the Naming Game (NG) a population of individuals agrees on the use of a simple convention (e.g. the name to give to an object) without resorting to any central coordination, but on the contrary exploiting only local interactions [7, 8]. It is possibly the simplest model in which the idea that language can be seen as a complex adaptive system [4] has been applied and challenged. This original seminal idea triggered a series of contributions along the same lines and many variants have been proposed over the years. It is particularly interesting to mention the work proposed in [9], that focuses on an imitation model which simulates how a common vocabulary is formed by agents imitating each other, either using a mere random strategy or a strategy in which imitation follows the majority (which implies non-local information for the agents). In the following we shall present a *minimal* [10] version of the NG which results in a drastic simplification of the model definition, while keeping the same overall phenomenology. Indeed, its simplicity has allowed for an extensive application of complex systems concepts and techniques to various aspects of its dynamics, ranging from the self-organizing global behaviors to the role of topology, that is unprecedented in the study of the emergence and evolution of language [10, 11]. For a review of statistical physics models related to language evolution we refer to [2].

2.1 The model

The NG is played by a population of N agents which play pairwise interactions in order to *negotiate* conventions, i.e., associations between forms and meanings, and it is able to describe the emergence of a global consensus among them. For the sake of simplicity the model does not take into account the possibility of homonymy, so that all meanings are independent and one can work with only one of them, without loss of generality. An example of such a game is that of a population that has to reach the consensus on the name (i.e., the form) to assign to an object (i.e., the meaning) exploiting only local interactions, and we will adopt this perspective in the following.

Each agent disposes of an internal inventory, in which an a priori unlimited number of words can be stored. As initial conditions we require all inventories to be empty. At each time step ($t = 1, 2, ..$), a pair of neighboring agents is chosen randomly, one playing as "speaker", the other as "hearer", and negotiate according to the following rules (see Fig. 1):

- the speaker selects randomly one of its words (or invents a new word if its inventory is empty) and conveys it to the hearer;
- if the hearer's inventory contains such a word, the two agents update their inventories so as to keep only the word involved in the interaction (*success*);
- otherwise, the hearer adds the word to those already stored in its inventory (*failure*).

Similar cooperative learning processes are present on the Web (e.g., in the case of collaborative tagging systems described in Sect. 4) and may be used in sensor networks [12].

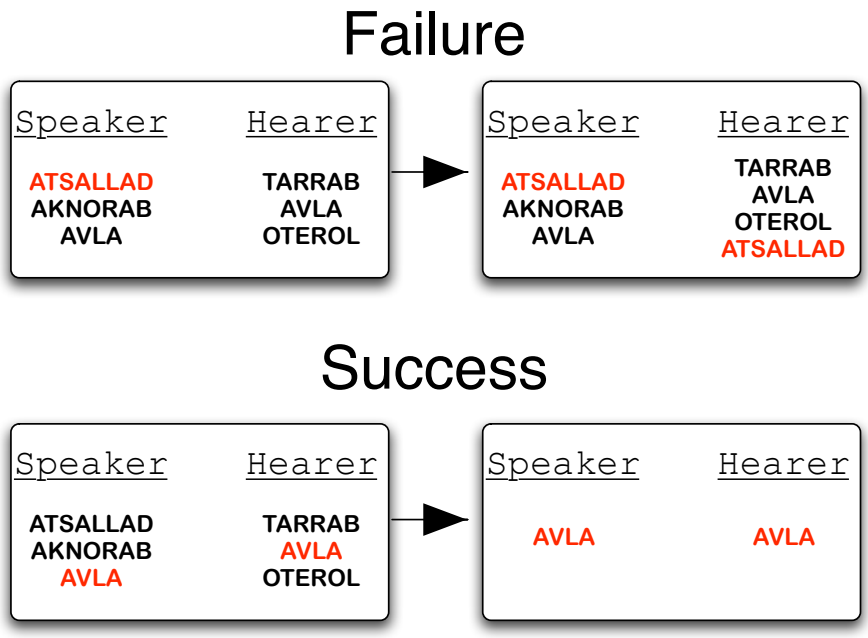

Figure 1: **Naming Game** Examples of the dynamics of the inventories in a failed (top) and a successful (bottom) game. The speaker selects the word highlighted. If the hearer does not possess that word he includes it in his inventory (top). Otherwise both agents erase their inventories only keeping the winning word (bottom).

With this scheme of interaction, the assumption of the absence of homonymy simply translates into assuring that each newly invented word had never appeared before in the population. Thus, single objects are independent (i.e., it is impossible that two agents use the same word for two different objects) and their number becomes a trivial parameter of the model. Indeed, one can treat an environment composed by an arbitrary number of objects by a simple controlled rescaling of the findings of the case with one object. For this reason, as we mentioned above, we concentrate on the presence of one single object, without loss of generality.

It is also interesting to note that the problem of homonymy has been studied in great detail in the context of evolutionary game theory and it has been shown [13] that languages with homonymy are evolutionary unstable. However it is obvious that homonymy is an essential aspect of human languages, while synonymy seems less relevant. The two authors solve this apparent paradox noting that if we think of "words in a context" homonymy almost disappears while synonymy acquires a much grater role. In the framework of the NG, homonymy is not always an unstable feature (see Sect. 3 and [14] for an example) and its survival depends in general on the size of the meaning and signal spaces [15].

This observation fits very well also with the implicitly assumed inferential model of learning, according to which we assume that agents are placed in a

common environment and they are able to point referents. So, after a failure, the speaker is able to point the named object (or referent) to the hearer which in its turn can assign the new name to it.

Another important remark concerns the random extraction of the word in the speaker's inventory. Most previously proposed models of Semiotic dynamics attempted to give a more detailed representation of the negotiation interaction assigning weights to the words in the inventories. In such models, the word with largest weight is automatically chosen by the speaker and communicated to the hearer. Success and failures are translated into updates of the weights: the weight of a word involved in a successful interaction is increased to the detriment of those of the others (with no deletion of words); a failure leads to the decrease of the weight of the word not understood by the hearer. An example of a model including weights dynamics can be found in [16] (and references therein). For the sake of simplicity the minimal NG described above avoids the use of weights. Indeed, these are apparently more realistic, but their presence is not essential for the emergence of a global collective behavior of the system [17].

2.2 Phenomenology

The non-equilibrium dynamics of the NG is characterized by three temporal regions: (1) initially the words are invented; (2) then they spread throughout the system inducing a reorganization process of the inventories; (3) this process eventually triggers the final convergence towards the global consensus (all agents possess the same unique word).

More precisely, the main quantities of interest which describe the system's evolution are [10]

- the total number $N_w(t)$ of words in the system at the time t (i.e., the total size of the memory);
- the number of different words $N_d(t)$ in the system at the time t;
- the average success rate $S(t)$, i.e., the probability, computed averaging over many simulation runs, that the chosen agent gets involved in a successful interaction at a given time t.

The consensus state is obtained when $N_d = 1$ and $N_w = N$ (so that $S = 1$), and the temporal evolution of the three main quantities is presented in Fig. 2 (circles). At the beginning, many disjoint pairs of agents interact, with empty initial inventories: they invent a large number of different words ($N/2$, on average) that start spreading throughout the system through failure events. Indeed, the number of words decreases only by means of successful interactions, which are limited in early stages by the very low overlap between the inventories. The number of different words N_d grows, rapidly reaching a maximum in a time that scales as $\mathcal{O}(N)$, and then saturates to a plateau where $N_d = N/2$, on average. The total number N_w of words, on the other hand, keeps growing after N_d has saturated, since the words continue to propagate throughout the system even if no new one is introduced. The peak of N_w scales as $\mathcal{O}(N^{1.5})$ [10], meaning that each agent stores $\mathcal{O}(N^{0.5})$ words. This peak occurs after the system has evolved for a time $t_{max} \sim \mathcal{O}(N^{1.5})$. In the subsequent dynamics, strong correlations between words and agents develop, driving the system to a final rather fast

convergence to the absorbing state in a time $t_{conv} \sim \mathcal{O}(N^{1.5})$. The S-shaped curve of the success rate in Fig. 2 summarizes the dynamics: initially, agents hardly understand each others ($S(t)$ is very low); then the inventories start to present significant overlaps, so that $S(t)$ increases until it reaches 1. It is worth noting that the established communication system is not only effective (agents can understand each others) but also efficient (no memory is wasted in the final state).

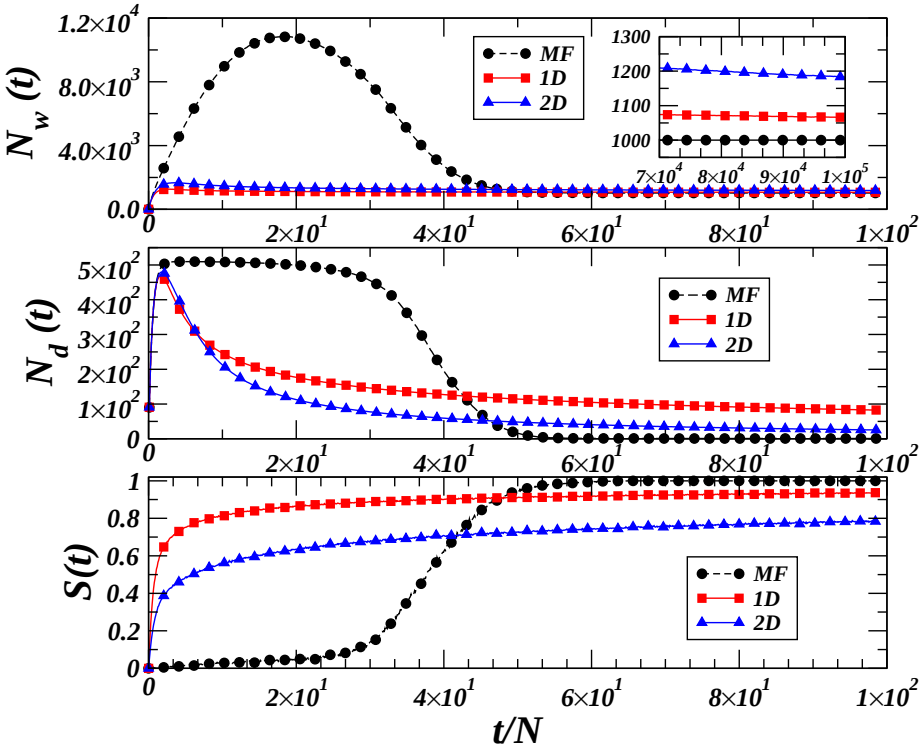

Figure 2: **Time evolution** of the total number of words N_w (top), of the number of different words N_d (center), and of the average success rate $S(t)$ (bottom), for a mean-field system (black circles) and low-dimensional lattices ($1D$, red squares and $2D$, blue triangles) with $N = 1024$ agents, averaged over 10^3 realizations. The inset in the top graph shows the very slow convergence in low-dimensional lattices.

2.3 Role of topology

In the NG at each time step two agents are *randomly* extracted. This means that each agent can in principle interact with anyone else, and the population is unstructured (i.e., we are in the mean-field case). However, the hypothesis that each agent can in principle talk to anybody else is quite unrealistic when we deal with large populations. The traditional remedy adopted in statistical mechanics is to embed agents in a regular quenched spatial structure, typically a regular $d-$dimensional lattice. The effects of such an underlying structure on the NG are remarkable [18]. Compared to the mean-field case, the model converges

	Mean-field	Lattices ($d \leq 4$)	Networks
Maximum memory	$N^{1.5}$	N	N
Convergence time	$N^{1.5}$	$N^{1+\frac{2}{d}}$	$N^{1.4\pm0.1}$

Table 1: Scaling with the system size N of the maximum number of words (memory) and time of convergence. Networks, thanks to the small-world property and the finite connectivity, ensure a trade-off between the fast convergence of mean-field topology and the small memory requirements of lattices.

very slowly, and the reason is related to the formation of many different local clusters of agents with the same unique word, that grow through a competition taking place at their boundaries (i.e., by coarsening dynamics, in the terms of statistical mechanics). Therefore the maximum memory per agent is now finite (while, as we have seen, it scales as $\sqrt{N}$ for the complete graph), but it can be shown that the consensus is reached in a time $t_{conv}/N \sim N^{2/d}$, i.e., slower than the mean-field case if $d \leq 4$.

A more realistic alternative to regular structures are given by complex networks. A network is, roughly speaking, an ensemble of nodes connected by links (or edges). Examples of such structures are ubiquitous, Internet and the World Wide Web being the most obvious. Moreover, recently, it has been found out that many more systems, ranging from the social to the biological and technological domains, can be described as networks [19, 20, 21]. Among the most peculiar features shared by most natural or artificial networks there are the "small world" property [22] and the scale free degree distribution [23]. The first is the name attributed to the evidence that the minimal hop distance between each pair of nodes scales logarithmically with the network's size instead of algebraically as in usual regular lattices. The second is the fact that, said degree k of a node the number of links which connect it to other nodes, the degree distribution $P(k)$ follows a power law $P(k) \sim k^{-\gamma}$, thus allowing for the presence of very few nodes with very high connectivity that in general play a central role in the structural and dynamical properties of the system.

The behavior of the NG on complex networks can be understood resorting to the results obtained in the most artificial cases of mean-field and low dimensional topologies [24, 25]. Indeed, the small-world property prevents the formation of compact clusters of agents sharing the same unique words, while the finite connectivity ensures finite memory requirements. Thus, the convergence time on these topologies is very close (except for logarithmic corrections) to the mean-field one, whereas the required memory stays finite (it does not depend on N only on the average degree). Table 1 summarizes the results we have presented.

Along with the global quantities we have studied so far, it is also interesting to investigate the microscopic activity patterns of single agents, and to study how they are affected by underlying topology [26]. In particular, in complex networks, simple properties of the degree distribution (namely the first two moments) turn out to affect dramatically the memory requirements of the agents, in a way that depends both on the general features of the considered network and on the connectivity of the single agents. Without entering in the mathematical details that allow to quantify precisely the impact of topology on agents activity [26], a simple look at Figure 3 allows to gain a qualitatively idea of the

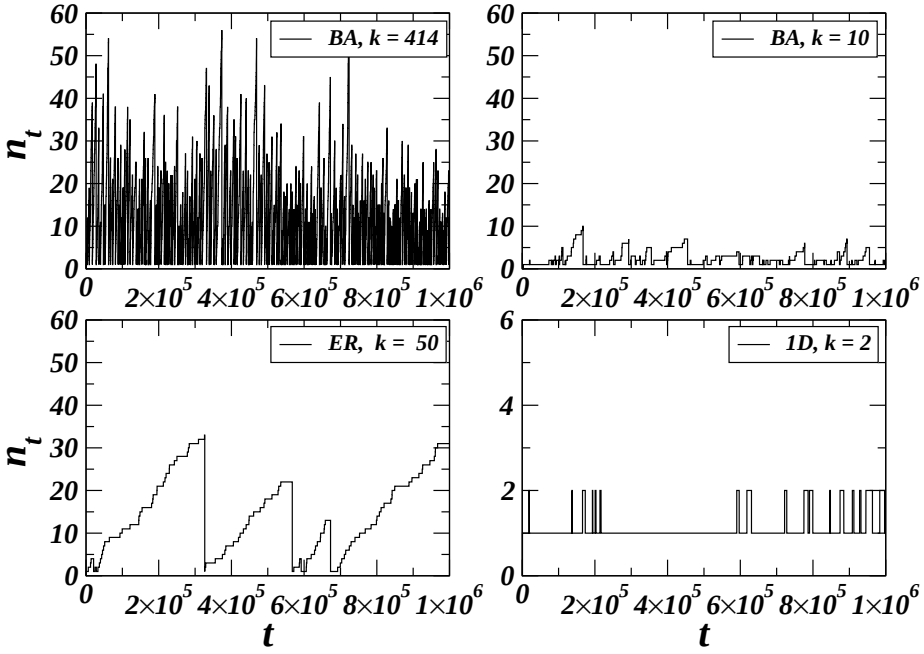

Figure 3: **Temporal series** of the inventory size of a single agent in different topologies. Top: Series from a Barabási-Albert (BA) network with $N = 10^4$ nodes and average degree $\langle k \rangle = 10$, for nodes of high degree (e.g. $k = 414$) and low degree (e.g. $k = 10$). Bottom: Series for nodes in Erdös-Rényi random graph ($N = 10^4$, $\langle k \rangle = 50$) and in a one-dimensional ring ($k = 2$).

relevance of the phenomenon. Here the time evolution of the inventory size of single agents is presented, and the role of connectivity patterns is evident. Top panels refers to an hub (left) and a less connected node (right) belonging to the same Barabási-Albert network [23] (an artificial graph with a power law degree distribution given by $P(k) \sim k^{-3}$, i.e., where the vast majority of nodes is poorly connected while few hubs have large degrees). The bottom left panel, on the other hand, concerns the activity of an average node on an homogeneous Erdös-Rényi random graph [27], where the degree distribution is peaked and all nodes have very similar connectivity patterns. The bottom right square, finally, presents the activity of an agent belonging to a population arranged on the nodes of a linear chain, whose inventory never exceeds the size of two words. In summary, the microscopic point of view not only supports and complements the study of global quantities, but also allows to point out deeper connections between the learning process of the agents (i.e., the dynamics of acquisition and deletion of words of a single agent) and the topological properties of the system.

3 The Category Game

In the past there have been many computational and mathematical studies addressing the learning procedures for form-meaning associations [28, 29]. From the point of view of methodology, the evolutionary scheme, based on the max-

imization of some fitness functions, has been extensively applied [30, 31]. Recent years, however, have shown that also the orthogonal approach of self-organization can be fruitfully exploited in multi-agent models for the emergence of language [32, 33, 10]. In this context, a community of language users is viewed as a complex dynamical system which has to develop a shared communication system [4, 34]. In this debate, a still open problem concerns the emergence of a small number of forms out of a diverging number of meanings. For example the few "basic color terms", present in natural languages, coarse-grain an almost infinite number of perceivable different colors [35, 36, 37].

Adopting the point of view of Semiotic dynamics [4], we shall show that an assembly of individuals with basic communication rules and without any external supervision may evolve an initially empty set of categories, achieving a non-trivial communication system characterized by a few linguistic categories [14]. To probe the hypothesis that cultural exchange is sufficient to this extent, individuals in this model are never replaced (unlike in evolutionary schemes [30, 31]), the only evolution occurring in their internal form-meaning association tables, i.e. their "mind". The individuals play elementary language games [6, 8] whose rules constitute the only knowledge initially shared by the population. They are also capable of perceiving analogical stimuli and communicating with each others [32, 33].

3.1 The model

The model involves a population of N individuals (or players), committed in the categorization of a single analogical perceptual channel, each stimulus being represented as a real-valued number ranging in the interval $[0, 1]$.

Here categorization is identified as a partition of the interval $[0, 1]$ in discrete sub-intervals, from now onwards denoted as perceptual categories. This approach can also be extended to categories with prototypes and fuzzy boundaries, for instance adding a weight structure upon it. Typical proposals in literature, such as prototypes with a weight function equal to the inverse of the distance from the prototype [33], are exactly equivalent to our "rigid boundaries" categories. Moreover, all the results of this experiment can be easily generalized to multi-dimensional perceptual channels, provided an appropriate definition of category domains is given. It should be kept in mind that the goal is to investigate why the continuum of perceivable meanings in the world is organized, in language, in a finite and small number of subsets with different names, with a no immediate (objective) cause for a given partition with respect to other infinite possibilities. Apart from the evident example of the partition of the continuous light spectrum in a small number of "basic color terms", this phenomenon is widespread in language: one can ask, for example, what objective differences allow to distinguish a cup from a glass; one can present a multi-dimensional continuum of objects able to "contain a liquid" (including also objects given as a prize), but a natural discontinuity between cups and glasses does not appear; our model, even reducing the phenomenon to the case of a 1-dimensional continuum, unveils a mechanism which can be easily extended to any kind of space, once it has been provided with a topology. The mechanism we propose for the discrete partition in linguistic subsets (categories) does not depend on the exact nature of this topology, which is of course a fundamental, yet different, matter of investigation.

Each individual has a dynamical inventory of form-meaning associations linking perceptual categories (meanings) to words (forms), representing their linguistic counterpart. Perceptual categories and words associated to them co-evolve dynamically through a sequence of elementary communication interactions, simply referred as games. All players are initialized with only the trivial perceptual category $[0, 1]$, with no name associated to it. At each time step a pair of individuals (one playing as speaker and the other as hearer) is selected and presented with a new "scene", i.e. a set of $M \geq 2$ objects (stimuli), denoted as $o_i \in [0, 1]$ with $i \in [1, M]$. The speaker discriminates the scene, adding new category boundaries to isolate the topic, then he names one object and the hearer tries to guess it.The word to name the object is chosen by the speaker among those associated to the category containing the object, with a preference for the one which has been succesfully used in the most recent game involving that category. A correct guess makes the game successful. Based on game's outcomes individuals may update their category boundaries and the inventory of the associated words: in a successful game both players erase competing words in the category containing the topic, keeping only the word used in that game; in failed games, the speaker points out the topic and the hearer proceeds to discriminate it, if necessary, and then adds the spoken word to its inventory for that category.

Fig. 4 describes two examples representing a failure (game 1) and a success (game 2), respectively. Two objects are presented to both players. The speaker selects the topic. In game 1 the speaker has to discriminate the chosen topic ("a" in this case) by creating a new boundary in his rightmost perceptual category at the position $(a + b)/2$. The two new categories inherit the words-inventory of the parent perceptual category (here the words "green" and "olive") along with a different brand new word each ("brown" and "blue"). Then the speaker browses the list of words associated to the perceptual category containing the topic. There are two possibilities: if a previous successful communication has occurred with this category, the last winning word is chosen; otherwise the last created word is selected. In the present example the speaker chooses the word "brown", and transmits it to the hearer. The outcome of the game is a failure since the hearer does not have the word "brown" in his inventory. The speaker unveils the topic, in a non-linguistic way (e.g. pointing at it), and the hearer adds the new word to the word inventory of the corresponding category. In game 2 the speaker chooses the topic "a", finds the topic already discriminated and verbalizes it using the word "green" (which, for example, may be the winning word in the last successful communication concerning that category). The hearer knows this word and therefore points correctly to the topic. This is a successful game: both the speaker and the hearer eliminate all competing words for the perceptual category containing the topic, leaving "green" only. In general when ambiguities are present (e.g. the hearer finds the verbalized word associated to more than one category containing an object), these are solved making an unbiased random choice.

The perceptive resolution power of the individuals limits their ability to distinguish objects/stimuli that are too close to each other in the perceptual space: in order to take this into account, we define a threshold d_{min}, inversely proportional to their resolution power [1]. In a given scene the M stimuli are chosen to be

[1] In psychology, d_{min} is equivalent to the so-called Just Noticeable Difference ("JND") or

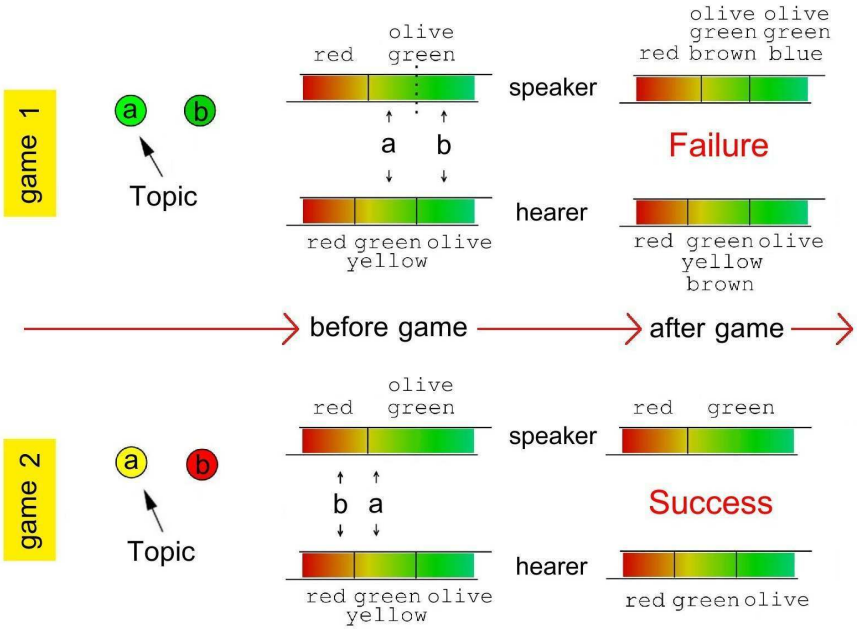

Figure 4: **Rules of the Category Game**. The two examples describe a failure (game 1) and a success (game 2), respectively (see text for the details).

at a distance larger than this threshold, i.e. $|o_i - o_j| > d_{min}$ for every pair (i, j). Nevertheless, objects presented in different games may be closer than d_{min}. The way stimuli are randomly chosen characterizes the kind of simulated environment: simulations will be presented both with a homogeneous environment (uniform distribution in $[0, 1]$) and more natural environments (e.g., without loss of generality, the distributions of the hue sampled from pictures portraying natural landscapes).

3.2 Phenomenology

A resume of the main results of the numerical experiment is given in Fig. 5. The evolution of the population presents two main stages: 1) a phase where players do not understand each other, followed by 2) a phase where communication has reached an averagely high success thanks to the emergence of a common language, still with evolving perceptual categories and a finite fraction of failures due to slightly unaligned categories and ambiguities. The first phase is marked by the growth and decline of synonymy, see Fig. 5a. Synonymy, in the context of the "naming game" (an individual object to be named), has been discussed in Sect. 2, where a similar evolution was observed and explained. All individuals, when necessary, create new words with zero probability of repetition: this leads to an initial growth of the vocabulary associated to each perceptual category. New words are spread through the population in later games and, whenever

Difference Limen ("DL")

a word is understood by both players, other competing words for the same category are forgotten [2]. This eventually leads to only one word per category. During the growth of the dictionary the success rate, see Fig. 5b, is very small. The subsequent reduction of the dictionary corresponds to a growing success rate which reaches its maximum value after synonymy has disappeared. In all numerical experiments the final success rate overcomes 80% and in most of them goes above 90%, weakly increasing with the final number of perceptual categories. Success is reached in a number of games per player of the order of 5×10^2, logarithmically depending on N, and it remains constant hereafter.

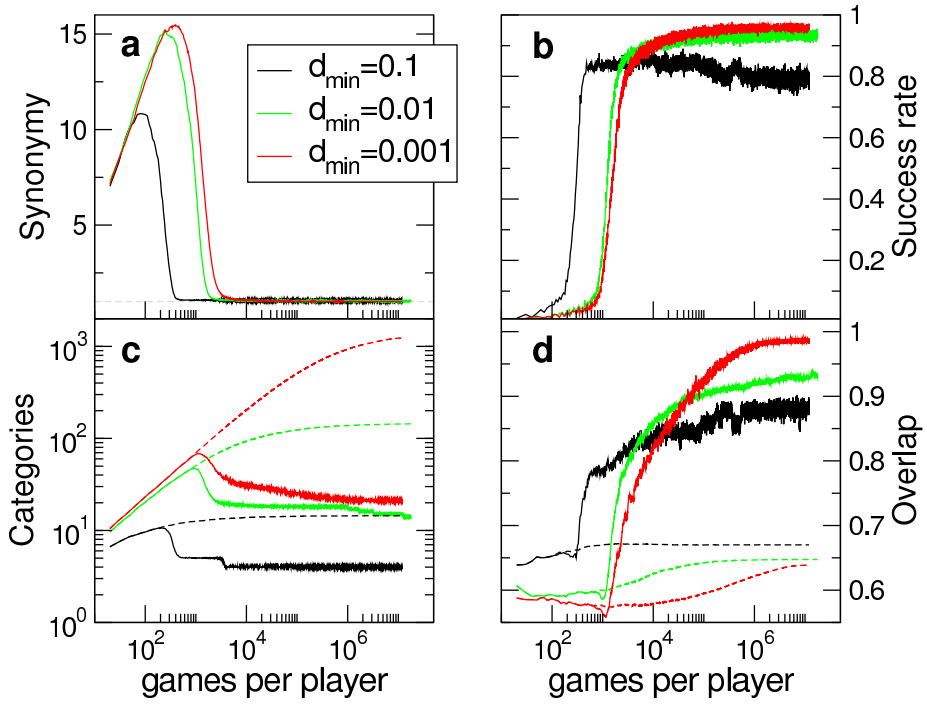

Figure 5: **Results of the simulations** with $N = 100$ and different values of d_{min}: a) Synonymy, i.e. average number of words per category; b) Success rate measured as the fraction of successful games in a sliding time windows games long; c) Average number of perceptual (dashed lines) and linguistic (solid lines) categories per individual; d) Averaged overlap, i.e. alignment among players, for perceptual (dashed curves) and linguistic (solid curves) categories.

The set of perceptual categories of each individual follows a somewhat different evolution (see dashed lines in Fig. 5c). The first step of each game is, in fact, the discrimination stage where the speaker (possibly followed by the hearer) may refine his category inventory in order to distinguish the topic from the other objects. The growth of the number of perceptual categories n_{perc} of each individual is limited by the resolution power: in a game two objects cannot appear at a distance smaller than d_{min} and therefore $n_{perc} < 2/d_{min}$.

[2] Extensions of this model can be devised in order to account for cases where words are not always erased, but instead they can became more "specialised", eventually yielding to the emergence of a hierarchy of category names.

The minimal distance also imposes a minimum number of categories $1/d_{min}$ that an individual must create before his discrimination process may stop. The average number of perceptual categories per individual, having passed $1/d_{min}$, grows sub-logarithmically and for many practical purposes it can be considered constant.

The success rate is expected to depend on the alignment of the category inventory among different individuals. The degree of alignment of category boundaries is measured by an *overlap function O*

$$O = 2\sum_{i<j} \frac{o_{ij}}{N(N-1)} \quad \text{with} \quad o_{ij} = \frac{2\sum_{c_i^j}(l_{c_i^j})^2}{\sum_{c_i}(l_{c_i})^2 + \sum_{c_j}(l_{c_j})^2}, \tag{1}$$

where l_c is the width of category c, c_i is one of the categories of the i-th player, and c_i^j is the generic category of the "intersection" set obtained considering all the boundaries of both players i and j. The function returns a value proportional to the degree of alignment of the category inventories, reaching its maximum unitary value when they exactly coincide. Its study, see dashed curves of Fig. 5d, shows that alignment grows with time and saturates to a value which is, typically, in between $60\% - 70\%$, i.e. quite smaller than the communicative success. This observation immediately poses a question: given such a strong misalignment among individuals, why is communication so effective?

The answer has to be found in the analysis of polysemy, i.e. the existence of two or more perceptual categories identified by the same unique word. Misalignment, in fact, induces a "word contagion" phenomenon. With a small but non zero probability, two individuals with similar, but not exactly equal, category boundaries, may play a game with a topic falling in a misalignment gap, as represented in Fig. 6a. In this way a word is copied to an adjacent perceptual category and, through a second occurrence of a similar event, may become the unique name of that category. Interfering events may occur in between: it is always possible, in fact, that a game is played with a topic object falling in the bulk of the category, where both players agree on its old name, therefore canceling the contagion. With respect to this cancelling probability, some gaps are too small and act as almost perfectly aligned boundaries, drastically reducing the probability of any further contagion. Thus, polysemy needs a two-step process to emerge, and a global self-organized agreement to become stable. On the other hand, polysemy guarantees communicative success: perceptual categories that are not perfectly aligned tend to have the same name, forming true linguistic categories, much better aligned among different individuals. The topmost curve of Fig. 5d, displays the overlap function measured considering only boundaries between categories with different names [3]: it is shown to reach a much higher value, even larger than 90%.

The appearance of linguistic categories is the evidence of a coordination of the population on a higher hierarchical level: a superior linguistic structure on top of the individual-dependent, finer, discrimination layer. The linguistic level emerges as totally self-organized and is the product of the (cultural) negotiation

[3] We define name of a perceptual category the word that an individual would choose, according to the rules of the model, to communicate about an object discriminated by that category (i.e. the last-winning word or the last created word). Of course, if there is a unique word associated with a category (which is most often the case after homonymy has almost disappeared), the definition above identifies that word as the name of the category.

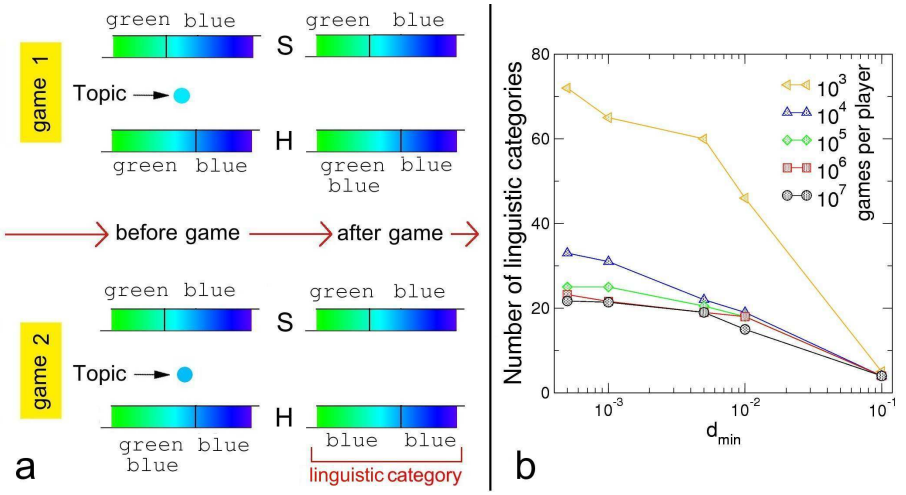

Figure 6: **Saturation in the number of linguistic categories**: a) A "word contagion" phenomenon occurs whenever the topic falls in a gap between two misaligned categories of two playing individuals. In the shown examples two individuals play two successive games. In game 1 the speaker (S) says "blue" and the hearer (H), unable to understand, adds "blue" as a possible word for his leftmost category; successively (game 2) the speaker repeats "blue" and the hearer learns this word as the definitive name for that perceptual category; both left and right perceptual categories of the hearer are now identified by the same name "blue" and they can be considered (for the purpose of communication) as a single linguistic category; b) Final number of linguistic categories as a function of d_{min} at different times, with $N = 100$. As the time increases the number of linguistic categories saturates. At large times, for small d_{min}, the number of linguistic categories becomes independent of d_{min} itself. Concerning size dependence, only a weak (logarithmic) dependence on N, not shown, is observed.

process among the individuals. The average number of linguistic categories per individual, n_{ling}, Fig. 5c (solid curves), grows together with n_{cat} during the first stage (where communicative success is still lacking), then decreases and stabilizes to a much lower value. Some configurations of both category layers, at a time such that the success rate has overcome 95%, are presented in Fig. 7, using different sets of external stimuli.

The analysis, resumed in Fig. 6b, of the dependence of n_{ling} on d_{min} for different times, makes these findings robust and, to our knowledge, unprecedented. As the resolution power is increased, i.e. as d_{min} is diminished, the asymptotic number of linguistic categories becomes less and less dependent upon d_{min} itself. Most importantly, even if any state with $n_{ling} > 1$ is not stable, we have the clear evidence of a saturation with time, in close resemblance with metastability in glassy systems [38, 39]. This observation allows to give a solution to the long-standing problem of explaining the finite (and small) number of linguistic categories n_{ling}. In previous pioneering approaches [32, 33] the number of linguistic categories n_{ling} was trivially constrained (with a small range of variability) by d_{min}, with a relation of the kind $n_{ling} \propto 1/d_{min}$, implying a divergence of n_{ling}

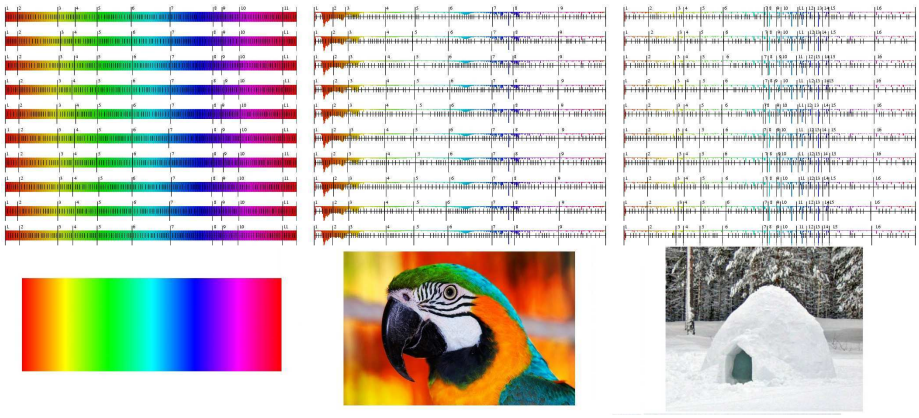

Figure 7: **Categories and the pressure of environment**. Inventories of 10 individuals randomly picked up in a population of $N = 100$ players, with $d_{min} = 0.01$, after 10^7 games. For each player the configuration of perceptual (small vertical lines) and linguistic (long vertical lines) category boundaries is superimposed to a colored histogram indicating the relative frequency of stimuli. The labels indicate the unique word associated to all perceptual categories forming each linguistic category. Three cases are presented: one with uniformly distributed stimuli (a) and two with stimuli randomly extracted from the hue distribution of natural pictures (b and c). One can appreciate the perfect agreement of category names, as well as the good alignment of linguistic category boundaries. Moreover, linguistic categories tend to be more refined in regions where stimuli are more frequent: an example of how the environment may influence the categorization process.

with the resolution power. In our model we have a clear indication of a finite n_{ling} even in the continuum limit, i.e. $d_{min} \to 0$, corresponding to an infinite resolution power.

3.3 Outlook on the Category Game

An extensive and systematic series of simulations has demonstrated that a simple negotiation scheme, based on memory and feedback, is sufficient to guarantee the emergence of a self-organized communication system which is able to discriminate objects in the world, requiring only a small set of words. Individuals alone are endowed with the ability of forming perceptual categories, while cultural interaction among them is responsible for the emergence and alignment of linguistic categories. The Category Game reproduces a typical feature of natural languages: despite a very high resolution power, the number of linguistic categories is very small. For instance, in many human languages, the number of "basic color terms" used to categorize colors usually amounts to about ten [35, 36, 37], in European languages it fluctuates between 6 and 12, depending on gender, level of education, and social class, while the light spectrum resolution power of our eyes is evidently much higher. Note that in the simulations one observes a reduction, with time, of the number of linguistic categories toward the final plateau. The experimental evidence [40], collected in empirical studies

on color categorization, of a growth of the number of categories from technologically less developed societies to more developed ones could be, in our opinion, an effect of the increased number N of players actively involved in the evolution of the communicative process. Finally we believe that these results could be important both from the point of view of language evolution theories, possibly leading to a quantitative comparison with real data [4] [41] and suggesting new experiments (e.g., different populations sizes and ages), and from the point view of applications (e.g., emergence of new communication systems in biological, social and technological contexts [42, 43]). The typical approach of physicists, to make contact between model and real data, would be to investigate the statistical properties of fluctuations (e.g. distribution of lengths of gaps between not perfectly aligned categories, or of number of different words in the population associated to some boundary region between adjacent categories). Since most of experimental data concern categorization of spaces of dimensionality larger than 1, an extension of the Category Game model to higher dimension is what is sought after in present and future researches.

4 Collaborative Tagging

The paradigm of collaborative tagging [44, 45] has been swiftly adopted and deployed in a wide range of systems, motivating a surge of interest in understanding their structure and evolution. Folksonomies have been known to exhibit striking statistical regularities and activity patterns [46, 43].

In this context, a natural topic for investigation is the vocabulary of tags that is used within a given system, and in particular its evolution over time, as new users, resources and tags come into play. Some insights in this direction are reported in [46] and [47], but a systematic attempt at characterizing vocabulary growth in collaborative tagging system is still lacking. Here we make a first step in that direction by analyzing a large-scale snapshot of *del.icio.us* and identifying a few stylized facts about the temporal evolution of tag vocabulary in a variety of contexts.

Ordinary vocabularies of words feature several interesting properties, and one of the most striking is related to their growth [48]. If one scans a text written in natural language and monitors the number of different words that have appeared as a function of the total number of words read, one realizes that this growth is described by a sub-linear law of growth, and often by a power-law behavior with an exponent smaller than one. It is thus tempting to investigate the same features in a folksonomy, regarded as a stream of time-ordered posts in a given context. How does the number of tags grow? Is the asymptotic number of tags finite? What is the rate of invention of new tags? Does their number eventually reach a plateau? Beyond the pure theoretical interest, these questions may be important for collaborative tagging and more generally for understanding the dynamics of tags in online social communities, where a deeper understanding of the temporal evolution of the system is important for both managing existing systems and designing new ones.

The outline of this section is as follows. Section 4.1 describes the experimental data we analyzed. Section 4.2 is devoted to the temporal evolution of the

[4] A collection of available experimental data can be found in *The World Color Survey*, http://www.icsi.berkeley.edu/wcs

global vocabulary, i.e. the growth of the number of different tags in the entire system, while the analysis of local vocabulary growth – the number of distinct tags used in the context of a given resource or user – is addressed in Section 4.3. In Section 4.4 we cast this work in a wider perspective and discuss some open questions.

4.1 Experimental data

Our analysis focuses on *del.icio.us* for several reasons: i) it was the first system to deploy the ideas and technologies of collaborative tagging, so it has acquired a paradigmatic character and it is the natural starting point for any quantitative study. ii) because of its popularity, it has a large community of active users and comprises a precious body of raw data on the structure and evolution of a folksonomy. iii) it is a *broad folksonomy* [49], i.e. single tagging events (posts) retain their identity and can be individually retrieved. This allows to define and measure the multiplicity (or frequency) of tags in a given context (for exampe, a resource or a user), providing a precious opportunity to probe social aspects in the tagging behavior of a community. Contrary to this, popular tagging systems falling in the *narrow folksonomy* class (*Flickr*, for example) are based on a different model of user interaction, where tags are mostly applied by the content creator, no notion of tag multiplicity is possible in the context of a single resource, and no access is given to the raw sequence of tagging events.

The basic unit of information in a collaborative tagging system is a `(user, resource, {tags})` triple, here referred to as "post". In *del.icio.us* (as well as in many other systems) a post also contains a timestamp indicating the physical time of the tagging event, so that the temporal ordering of posts can be preserved and the dynamical evolution of the system over time can be reconstructed and investigated.

The dataset used for the present analysis consists of approximately $5 \cdot 10^6$ posts, comprising about 650000 users, $1.9 \cdot 10^6$ resources and $2.5 \cdot 10^6$ distinct tags, and covering almost 3 years of user activity, from early 2004 up to November 2006. For the present study, a time-ordered sequence of posts was built and converted to a time-ordered table of tag assignments (TAS), by mapping each post of the form `(user, resource, {tag1, tag2, ... })` into adjacent assignments of the form `(tag1, user, resource), (tag2, user, resource),` ... , one for each tag in the post. Such a table, and selections of it, were used as the base for analysis described in the following.

4.2 Global vocabulary growth

We begin by studying the evolution over time of the size of the global "tag vocabulary", i.e. the total number of different tags that are present in the folksonomy. As a function of physical time (inset of Fig. 8) the growth of the global vocabulary is rather featureless, and reflects the huge growth of *del.icio.us* over the past 3 years. The fact that the system grew up in size so fast, indeed, makes physical time unsuitable to study the temporal evolution of *del.icio.us*, because a large fraction of the total activity is compressed in the final part of its temporal history. Physical time is in many respects "external" to the system, and a much better notion of "time" can be defined in terms of quantities that are intrinsic to the system itself. As mentioned above, we start our analysis

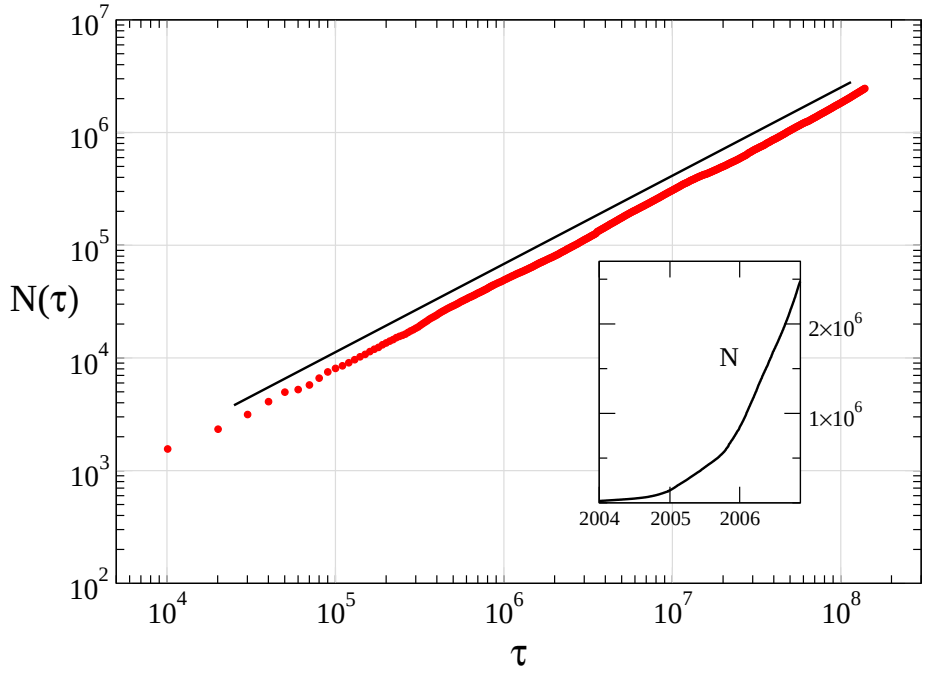

Figure 8: **Temporal evolution of the total number of distinct tags in *del.icio.us*** As a function of the intrinsic time τ (see main text), the number $N(\tau)$ of distinct tags (red dots) increases closely following a power-law (straight line in a log-log plot) across the entire history of the system. The solid black line, provided as an aid for the eye, corresponds to a power-law with exponent $\gamma \simeq 0.8$. The inset shows the number N of distinct tags as function of physical time, spanning almost 3 years of growth and six orders of magnitude in vocabulary size. The main graph and the inset refer to the same interval of physical time.

from a time-ordered table of tag assignments. For the system as a whole, we can define an "intrinsic time" τ as the index of a tag assignment into such a table, so that τ runs from 1 to the number of total tags assignments, i.e. the sum of the number of tags in all posts (about $1.4 \cdot 10^8$ in our case). For each post added to the system, this "clock" τ increases by a number of ticks equal to the number of tags in that post.

Fig. 8 shows the total number of distinct tags $N(\tau)$ present in the system at time τ, as a function of τ. In terms of this intrinsic time, a remarkably clean power-law behavior (straight line on a log-log plot) can be observed throughout the full history of the system. This is even more interesting because the data shown in Fig. 8 span a time interval covering almost the entire history of *del.icio.us*: the power-law trend emerges already at the very beginning and is obeyed all the way to present times, as the number of active users and that of bookmarked resources dramatically increase over time. It's worth noticing the following points:

- The number N of distinct tags present in the system does not appear to level off towards a steady-state plateau. This is not surprising in its own merit because *del.icio.us* is an open-ended system and new users and

resources are a source of continuous novelty for the tags comprised by the folksonomy.

- The power-law growth followed by $N(\tau)$ is of the form $N(\tau) \sim \tau^{\gamma}$, with $\gamma < 1$. The black line in Fig. 8 corresponds to $\gamma \simeq 0.8$.
- The rate at which new tags appear at time τ scales as $dN(\tau)/d\tau \sim \tau^{\gamma-1}$. That is, new tags – as a function of the intrinsic time τ – appear less and less frequently, with the invention rate of new tags monotonically decreasing towards zero. The approach to zero is however so slow that the cumulated number of tags, asymptotically, does not converge to a constant value but is unbounded — assuming the observed trend stays valid.

It is remarkable that the above statistical regularities hold throughout the history of *del.icio.us*, while the system undergoes a huge change in the size of its user base, the number of bookmarked resources, several changes in the user interface are made, tag suggestion is introduced, and so on.

The above observations constitute the core facts of the present study, and in the following we will shift from the global view of the system to a local one, to see whether these facts stay valid, and to deepen our analysis.

4.2.1 Sub-linearity in vocabulary growth

The sub-linear growth reported here is not a newly observed phenomenon. When dealing with the evolution of the number of attributes pertaining to some collection of objects, this sub-linear growth is generally referred to as *Heaps' law* [48]. As an example, sub-linear behavior has been observed in the growth of vocabulary size in texts, i.e. in the number of different words in a text as a function of the total number of words observed while scanning through it. For the case of English corpora, vocabulary growth exponents in the range $0.4 < \gamma < 0.6$ have been reported [50]. The vocabulary size of the Thai subset of WWW internet web pages has also been found to obey a sub-linear power-law behavior with exponent $\gamma \approx 0.5$ [51]. In contrast, the exponent we observe here is comparatively high. Attempts to explain the power-law behaviors of vocabulary growth in terms of the measured Zipf's frequency-rank distribution of words can be found in literature [52], as well as ad-hoc modifications of simple stochastic models [53].

It is important to remark that, at odds with texts, no grammatical structure is embedded in tags. Moreover, the words used in folksonomies are mainly nouns or, in general, synthetic descriptions of categories [54]. In this sense, the only linguistic mechanisms that could be responsible for a sub-linear growth is a possibly hierarchical organization of tags induced by semantics. Another important difference is that in written texts the number of authors is usually limited, while the number of users contributing to the tag vocabulary of *del.icio.us* is large and growing in time.

4.3 Local vocabulary growth

We will now shift our focus to a local scope, moving from a global view of tag vocabulary to a more fine-grained one, dealing with the restricted contexts of single resources and users. Specifically, we will investigate how the number of

different tags associated with a given resource (or user) grows as a function of an intrinsic time. The notion of time we adopt in the following is the same we employed for the global analysis of Section 4.2, except that in this case it is restricted to the context of a single resource or user: given a resource (or a user), we select from the global, time-ordered TAS table only those tag assignments that involve that resource (or user). We define the intrinsic time τ as the index into the selection, i.e. the cumulated number of tags associated with that resource (or user). Thus our notion of time is resource-dependent (or user-dependent), and τ naturally measures metadata accumulation in the specific semantic context of a single resource or user.

4.3.1 Resource-specific vocabularies

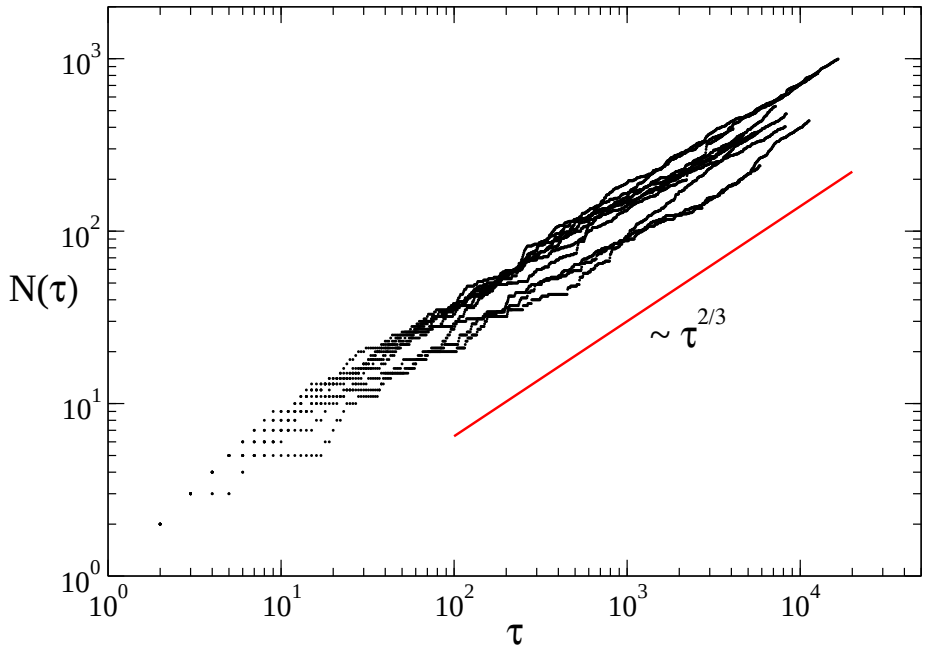

Figure 9: **Vocabulary growth for single resources** For 10 different resources in *del.icio.us*, the number of distinct tags $N(\tau)$ associated with them is plotted as a function of the intrinsic time τ pertaining to each resource. While single resources display a somehow noisy evolution, an overall power-law behavior governing the vocabulary growth is apparent, with an exponent $\gamma \simeq 2/3$ (red line).

The large number of users tagging each resource make the statistical features of resources quite similar, as long as they are popular enough. This can be shown, for instance, by looking at the vocabulary growth in the context of a single resource. In Fig. 9 we consider 10 different popular resources and we plot the number $N(\tau)$ of distinct tags associated with them by the entire user community, as a function of the intrinsic (resource-specific) time τ, the total number of tags assigned. The resources are chosen among the 1000 top-bookmarked resources in the system, starting from rank 100 and decreasing

at intervals of 100. While the vocabulary growth exhibits a somehow noisy temporal evolution, the general trend of growth appears to be compatible with an algebraic law of growth, a power-law with an exponent close to 2/3. This is a striking regularity, valid for very different resources across the system. Also, at this level of detail, no systematic dependence on the popularity of a resource can be detected. The local exponent of growth is smaller than the global one (Fig. 8) and the relation between the two may be linked to the statistical properties of tag co-occurrence, and might ultimately provide insights into the semantic structure of folksonomies.

To better probe the similarity of growth behaviors for different resources, we defined a rescaled growth curve, where both the intrinsic time τ and the final number of distinct tags $N(\tau_{max})$ are divided by their final values, τ_{max} and $N(\tau_{max})$, respectively. In this way, the curves for different resource can be easily plotted on the same graph. As shown in Fig. 10, all the rescaled curves lie between two limit power-laws, $(\tau/\tau_{max})^1$ and $(\tau/\tau_{max})^{1/2}$. More importantly,

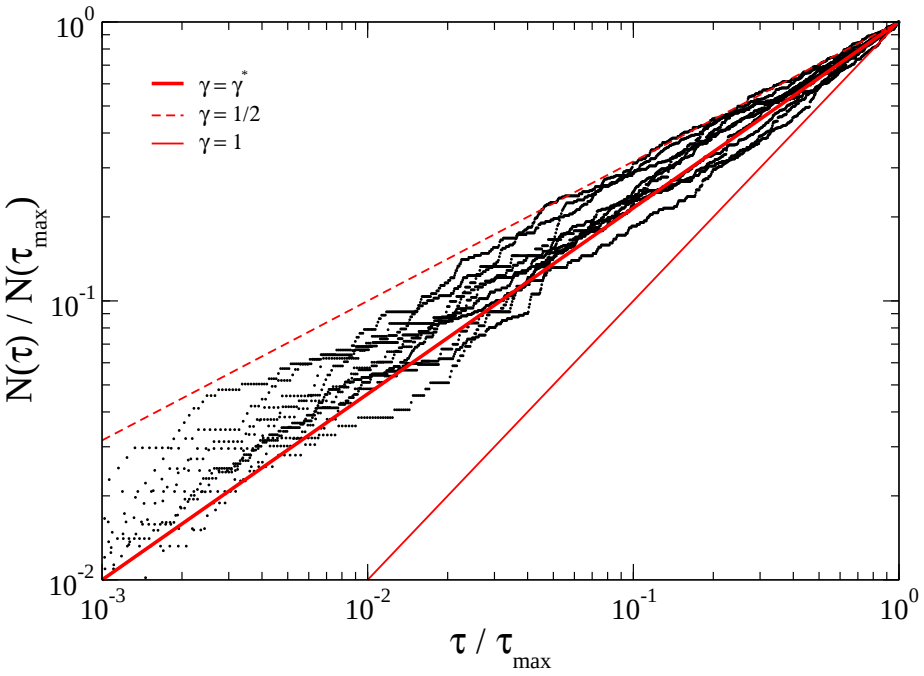

Figure 10: **Rescaled vocabulary growth** The curves of Fig. 9 were rescaled by dividing both the intrinsic time τ and the number of distinct tags $N(\tau)$ by their final (resource-specific) values $\tau_{\max}$ and $N(\tau_{\max}$, respectively. After rescaling, all curves lie approximately along the "universal" $(\tau/\tau_{\max})^{2/3}$ line (thick red line). On approaching the common endpoint, the slope of all curves appear to lie in the 0.5-1 range (dashed line and thin red line, see also Fig. 11).

4.3.2 Distribution of growth exponents

In order to make a more quantitative measure over a broader set of resources, we implement the following unsupervised procedure for characterizing the growth of local tag vocabularies: for each resource we measure an effective exponent γ that approximates the rescaled vocabulary growth with a power-law $(\tau/\tau_{max})^{\gamma}$. The simplest way to do this is to compute γ as $\gamma = \log(N(\tau_{max}))/\log(\tau_{max})$. Fig. 11 shows the probability distribution of the resulting values of γ, measured for three groups of resources. In particular, the red curve in Fig. 11 displays the distribution of γ values for the 1000 top ranked (most bookmarked) resources in *del.icio.us*. The distribution is well approximated by a rather narrow Gaussian distribution, whose average value is $\gamma^* \simeq 0.7$. This seems to confirm the idea (Fig. 10) that there is a well-defined exponent of growth governing the temporal evolution of popular resources. Moreover, the vocabulary growth of popular resources appears slower than the system-wide vocabulary growth of Fig. 8.

On computing the distribution $P(\gamma)$ for less and less popular resources (black and blue curves), the distribution gets broader and its peak shifts towards higher values of γ, indicating that the growth behavior is becoming more and more linear. This crossover from sub-linear to linear growth for resources bookmarked by just a few users is expected and corresponds to a sort of "priming" effect for the resource: the first few users who bookmark it build the "core" tag vocabulary for the resource, and since only a few posts are present at that time, most tags are new and the size of the vocabulary grows linearly with the total number of tags τ as well as with the number of posts associated with the resource. As more and more users bookmark the resource, correlations and social effects come into play and the law of growth crosses over from the linear to the "universal" sub-linear behavior reported above.

To make contact between local vocabulary growth in the context of a single resource and vocabulary growth in the context of a single user, we repeat the above analysis for the 1000 most active users in *del.icio.us* (as measured by the number of resources they bookmarked). The resulting probability distribution $P(\gamma)$ is shown in Fig. 12 and is qualitatively similar to the ones of Fig. 11. In particular, we notice that the peak of $P(\gamma)$ is compatible with the value γ^* observed for the top-ranked resources.

We would like to remark that the huge variability of vocabularies, at the level of single users and resources, is not in contrast with very regular – and simple – features at the global level. On the contrary, the emergence of regularity from the uncoordinated activity of users is the hallmark of complexity and indicates that tools and concepts from complex system science may prove valuable for understanding the structure and dynamics of folksonomies.

4.4 Outlook on Collaborative Tagging

Analyzing the growth dynamics of local vocabularies, associated with a given resource or user, may provide insights into the relationship between the behavior of individual users and vocabulary growth at the system – or community – level, as well as insights into the process of invention of new tags. For popular resources in *del.icio.us* we report a sub-linear growth with exponents sharply peaked around a characteristic value (slightly different from the global one), while for less popular resources we observe exponent values slowly shifting to-

wards 1. The sub-linear growth observed at the local level cannot be explained as a mere reflection of the growth in the number of users, (which, for resources, is linear in the intrinsic time) nor as an increase in the average number of tags per post (which has a rather stable characteristic value).

These observations point out that sub-linear dictionary growth is a genuine non-trivial feature of the system and open several problems. Is sub-linear growth at the global level (or at the local level) related to correlations among users' activity? Does the growth observed in the context of a single user reflect a collective/cooperative phenomenon, or is it just mirroring the complex cognitive processes (incorporating semantics) at the level of that individual user? Is the difference between local and global exponents relevant, and if so, what kind of information about the structure of tag space is it disclosing? What are the key elements in the user-system interaction that lead to the observed behaviors?

5 Conclusions

In this paper we have illustrated the point of view of statistical physics on the longstanding problem concerning the emergence and evolution of language. We introduced in particular the main tools and methods proposed so far by describing several concrete examples corresponding to the early stages in the emergence of language: e.g., the formation of a shared lexicon and the establishment of a common set of linguistic categories. Though promising, these studies did not yet face the hardest problems in linguistics, namely the emergence of syntax and grammar. Currently new studies are ongoing focusing on the emergence of higher forms of agreement, e.g., compositionality, syntactic or grammatical structures. It is clear how it would be highly important to cast a theoretical framework where all these problems could be defined, formalized and solved. In this perspective a crucial factor will be most likely represented by the availability of large sets of empirical quantitative data. The joint interdisciplinary activity should then include systematic campaigns of data gathering as well as the devising of new experimental setups for a continuous monitoring of linguistic features. From this point of view the Web may be of great help, both as a platform to perform controlled online social experiments, and as a repository of empirical data on large-scale phenomena. It is only in this way that a virtuous cycle involving data collection, data analysis, modeling and predictions could be triggered, giving rise to an ever more rigorous and focused approach to language.

It is worth stressing how the contribution physicists could give should not be considered in any way as alternative to more traditional approaches. We rather think that it would be crucial to foster the interactions across the different disciplines cooperating with linguistics, by promoting scientific activities with concrete mutual exchanges among all the interested scientists. This would help both in identifying the problems and sharpening the focus, as well as in devising the most suitable theoretical concepts and tools to approach the research.

6 Acknowledgments

A. Baronchelli acknowledges support from the DURSI, Generalitat de Catalunya (Spain) and from Spanish MEC (FEDER) through project No: FIS 2007-66485-CO2-01. This work was partially supported by the EU under contract IST-1940 (ECAgents) and IST-34721 (TAGora). The ECAgents and TAGora projects are funded by the Future and Emerging Technologies program (IST-FET) of the European Commission.

References

[1] V. Loreto and L. Steels. Social dynamics: Emergence of language. *Nature Physics*, 3:758–760, 2007.

[2] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. 2007, arXiv:0710.3256v1.

[3] M. Buchanan. *The social atom*. Bloomsbury, New York, NY, USA, 2007.

[4] L. Steels. Language as a complex adaptive system. In M. Schoenauer, editor, *Proceedings of PPSN VI*, Lecture Notes in Computer Science, Berlin (Germany), 2000. Springer-Verlag.

[5] L. Wittgenstein. *Philosophische Untersuchungen*. Suhrkamp Verlag, Frankfurt am Main, Germany, 1953.

[6] L. Wittgenstein. *Philosophical Investigations. (Translated by Anscombe, G.E.M.)*. Basil Blackwell, Oxford, UK, 1953.

[7] L. Steels. A self-organizing spatial vocabulary. *Artif. Life*, 2(3):319–332, 1995.

[8] L. Steels. Self-organizing vocabularies. In C. Langton and T. Shimohara, editors, *Artificial Life V: Proceeding of the Fifth International Workshop on the Synthesis and Simulation of Living Systems*, pages 179–184, Cambridge, MA, USA, 1996. The MIT Press.

[9] J. Ke, J. Minett, C-P. Au, and W. Wang. Self-organization and selection in the emergence of vocabulary. *Complexity*, 7(3):41–54, 2002.

[10] A. Baronchelli, M. Felici, E. Caglioti, V. Loreto, and L. Steels. Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics*, P06014, 2006.

[11] A. Baronchelli, L. Dall'Asta, A. Barrat, and V. Loreto. Bootstrapping communication in language games: Strategy, topology and all that. In Angelo Cangelosi, Andrew D M Smith, and Kenny Smith, editors, *The Evolution of Language: Proceedings of the 6th International Conference (EVOLANG6)*. World Scientific Publishing Company, 2006.

[12] Q. Lu, G. Korniss, and B.K. Szymanski. Naming games in spatially-embedded random networks. In *AAAI Fall Symposium Series: Interaction and Emergent Phenomena in Societies of Agents*, 2006.

[13] N. Komarova and P. Niyogi. Optimizing the mutual intelligibility of linguistic agents in a shared world. *Artif. Intell.*, 154(1-2):1–42, 2004.

[14] A. Puglisi, A. Baronchelli, and V. Loreto. Cultural route to the emergence of linguistic categories. March 2007, arxiv:physics/0703164.

[15] G. Gosti. Role of the homonymy in the naming game. Undergraduate thesis, "Sapienza" Univ. of Rome, 2007, 2007.

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

[17] A. Baronchelli. *Statistical mechanics approach to language games*. PhD thesis, Università Degli Studi di Roma "La Sapienza", January 2007.

[18] A. Baronchelli, L. Dall'Asta, A. Barrat, and V. Loreto. Topology-induced coarsening in language games. *Phys. Rev. E*, 73(1):015102, 2006.

[19] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Review of Modern Physics*, 74:559–564, 2002.

[20] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, Cambridge (USA), 2004.

[21] G. Caldarelli. *Scale Free networks*. Oxford University Press, Oxford (UK), 2007.

[22] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small world' networks. *Nature*, 393:440, 1998.

[23] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[24] L. Dall'Asta, A. Baronchelli, A. Barrat, and V. Loreto. Agreement dynamics on small-world networks. *Europhys. Lett.*, 73(6):969–975, 2006.

[25] L. Dall'Asta, A. Baronchelli, A. Barrat, and V. Loreto. Non-equilibrium dynamics of language games on complex networks. *Phys. Rev. E*, 74:036105, 2006.

[26] L. Dall'Asta and A. Baronchelli. Microscopic activity patterns in the naming game. *J. Phys. A: Math. Gen.*, 39:14851–14867, 2006.

[27] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[28] J. Hurford. Biological evolution of the Saussurean sign as a component of the language acquisition device in linguistic evolution. *Lingua*, 77:187, 1989.

[29] T. Briscoe, editor. *Linguistic evolution through language acquisition*. Cambridge University Press, Cambridge (UK), 2002.

[30] M. A. Nowak and D. C. Krakauer. The evolution of language. *Proc. Natl. Acad. Sci. USA*, 98:13189, 1999.

[31] M. A. Nowak, J. B. Plotkin, and D. C. Krakauer. The evolutionary language game. *Jour. Theor. Bio.*, 200:147, 1999.

[32] L. Steels and T. Belpaeme. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28:469–529, 2005.

[33] T. Belpaeme and J. Bleys. Explaining universal color categories through a constrained acquisition process. *Adaptive Behavior*, 13:293–310, 2005.

[34] N. L. Komarova. Population dynamics of human language: A complex system. In *Frontiers of engineering: reports on leading-edge engineering from the 2005 symposium*, pages 89–98, 2006.

[35] B. Berlin and P. Kay. *Basic color terms: Their universality and evolution*. University of California Press, 1969 (reprinted 1991).

[36] D. T. Lindsey and A. M. Brown. Universality of color names. *Proc. Natl. Acad. Sci. US*, 103:16608, 2006.

[37] B. A. C. Saunders and J. van Brakel. Are there nontrivial constraints on colour categorization? *Behavioral and Brain Sciences*, 20:167, 1997.

[38] M. M. Mezard, G. Parisi, and M. A. Virasoro. *Spin glass theory and beyond*. World Scientific, New Jersey (US), 1987.

[39] P. G. Debenedetti and F. H. Stillinger. Supercooled liquids and the glass transition. *Nature*, 410:259, 2001.

[40] P. Kay and L. Maffi. Color appearance and the emergence and evolution of color lexicons. *American Anthropologist*, 101:743, 1999.

[41] R. Selten and M. Warglien. The emergence of simple languages in an experimental coordination game. *Proc. nat. Acad. Sci. USA*, 104:7361, 2007.

[42] L. Steels. Semiotic dynamics for embodied agents. *IEEE Intelligent Systems*, 21:32, 2006.

[43] C. Cattuto, V. Loreto, and L. Pietronero. Semiotic dynamics and collaborative tagging. *Proc. Natl. Acad. Sci. USA*, 104:1461–1464, 2007.

[44] A. Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata, December 2004. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

[45] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11, Apr 2005. 10.1045/april2005-hammond.

[46] S. Golder and B.A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006, cs.DL/0508082.

[47] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM Press.

[48] H. S. Heaps. *Information Retrieval-Computational and Theoretical Aspects*. Academic Press, 1978.

[49] T. Vander Wal. Explaning and showing broad and narrow folksonomies, 2005.

[50] *Overview of the Third Text REtrieval Conference*. NIST Special Publication 500-207, 1995. NIST Special Publication 500-207.

[51] S. Sanguanpong, S. Warangrit, and K. Kohtarsa. Facts about the thai web, 2000.

[52] D. C. van Leijenhorst and Th. P. van der Weide. A formal derivation of heaps' law. *Information Sciences*, 170(2-4):263–272, 2005.

[53] D.H. Zanette and M.A. Montemurro. Dynamics of text generation with realistic zipf's distribution. *Journal of Quantitative Linguistics*, 12(1):29–40, 2005.

[54] C. Veres. The language of folksonomies: What tags reveal about user classification. In Christian Kop, Günther Fliedl, Heinrich C. Mayr, and Elisabeth Métais, editors, *NLDB*, volume 3999 of *Lecture Notes in Computer Science*, pages 58–69. Springer, 2006.

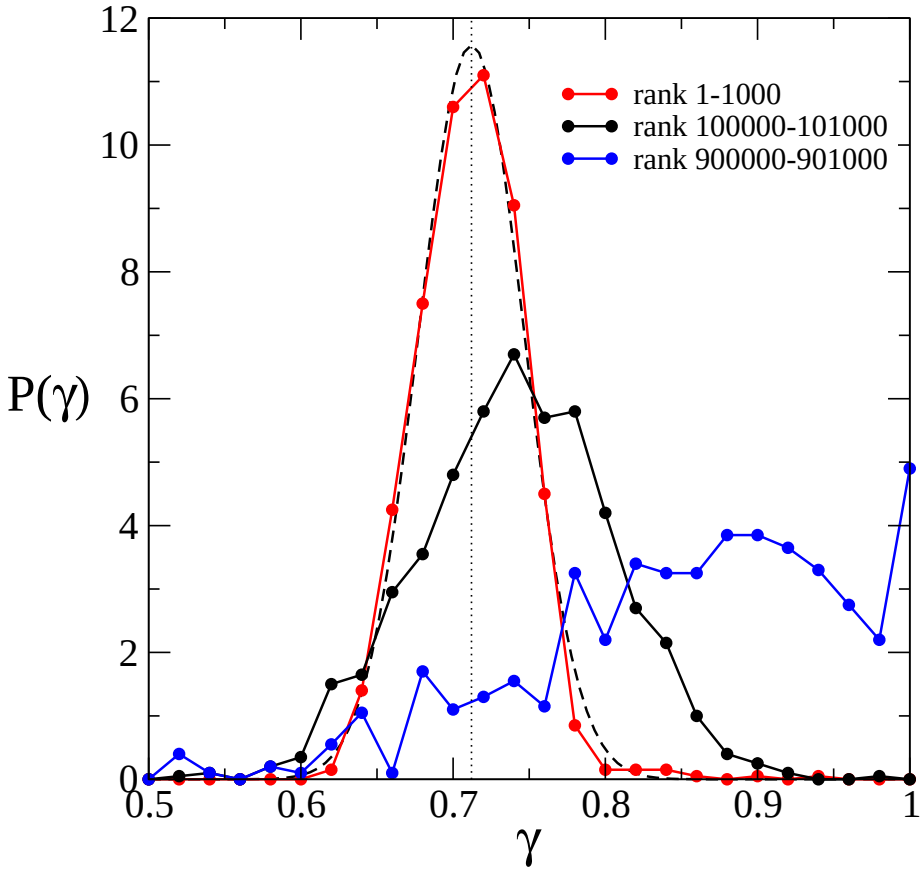

Figure 11: **Probability distribution of the vocabulary growth exponent** γ for resources, as a function of their rank. The red curve is the normalized probability distribution $P(\gamma)$ for the 1000 top-ranked (most bookmarked) resources in *del.icio.us*. It appears to be sharply peaked at a characteristic value $\gamma^* \simeq 0.71$ (vertical line) and can be closely fitted with a Gaussian (dashed line). This indicates that highly bookmarked resources share a characteristic law of growth, as already pointed out in Fig. 10. On computing the distribution $P(\gamma)$ for less and less popular resources (black curve and blue curve), the peak shifts towards higher values of γ and the growth behavior becomes more and more linear. The typical number of users who have bookmarked the resources used in this analysis is approximately a few thousands for the red curve, a few hundreds for the black curve, and just a few users for the blue one.

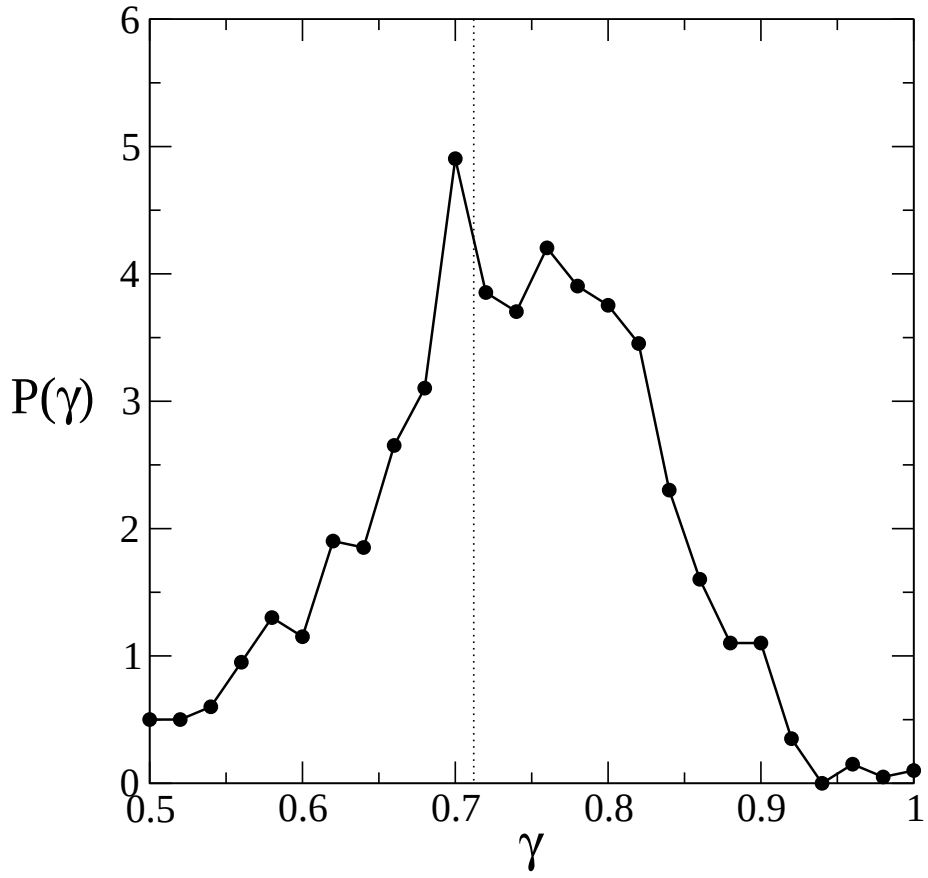

Figure 12: **Probability distribution of the vocabulary growth exponent** $\gamma$for user vocabularies. The distribution $P(\gamma)$ was computed for the 1000 most active users in *del.icio.us*. Similarly to Fig. 11, it appears peaked around a characteristic value close to the same observed for top-ranked resources (vertical line, same as in Fig. 11).