TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 8 (2016) 469–480

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12193

On the Emergence of Syntactic Structures: Quantifying and Modeling Duality of Patterning

Vittorio Loreto,[a,b,c] Pietro Gravino,[a,d] Vito D. P. Servedio,[a] Francesca Tria[b]

[a]*Physics Department, Sapienza University of Rome*
[b]*Institute for Scientific Interchange ISI*
[c]*SONY Computer Science Lab*
[d]*Physics Department, University of Bologna*

Received 3 December 2013; received in revised form 29 January 2015; accepted 29 January 2015

Abstract

The complex organization of syntax in hierarchical structures is one of the core design features of human language. Duality of patterning refers, for instance, to the organization of the meaningful elements in a language at two distinct levels: a combinatorial level, where meaningless forms are combined into meaningful forms; and a compositional level, where meaningful forms are composed into larger lexical units. The question remains wide open regarding how such structures could have emerged. The aim of this paper is to address these two aspects in a self-consistent way. First, we introduce suitable measures to quantify the level of *combinatoriality* and *compositionality* in a language, and we present a framework to estimate these observables in human natural languages. Second, we show that a recently introduced multi-agent modeling scheme, namely the *Blending Game*, provides a mathematical framework to address the problem of how a population of individuals can bootstrap combinatoriality and compositionality. Theoretical predictions based on this model turn out to be in good agreement with empirical data. It is remarkable that the two sides of duality of patterning emerge simultaneously as a consequence of a pure cultural dynamics in a simulated environment that contains meaningful relations, provided a simple constraint on message transmission fidelity is also considered.

Keywords: Combinatoriality; Compositionality; Language dynamics; Quantitative linguistics

1. Introduction

In a seminal paper, Charles Hockett (1960) identified duality of patterning as one of the core design features of human language. A language exhibits duality of patterning

Correspondence should be sent to Pietro Gravino, Physics Department, Sapienza University of Rome, Piazzale Aldo Moro 2, 00185 Rome, Italy. E-mail: vittorio.loreto@gmail.com

when it is organized at two distinct levels. At a first level, meaningless forms (typically referred to as phonemes) are combined into meaningful units (henceforth this property will be referred to as *combinatoriality*). For example, the English forms /k/, /a/, and /t/ are combined in different ways to obtain the three words /kat/, /akt/, and /tak/ (respectively written *cat*, *act*, and *tack*). Because the individual forms in them are meaningless, these words have no relation in meaning in spite of being made up of the same forms. This is a very important property, thanks to which all of the many words of the English lexicon can be obtained by relatively simple combinations of few dozens of phonemes. If phonemes had individual meaning, this degree of compactness would be hardly possible since phonemes would tend to proliferate in order to accomplish a holistic situation (one phoneme per each meaning). At a second level, meaningful units (typically referred to as morphemes) are composed into larger units, the meaning of which is related to the individual meaning of the composing units (henceforth this property will be referred to as *compositionality*). For example, the meaning of the word *boyfriend* is related to the meaning of the words *boy* and *friend* that compose it. The compositional level includes syntax as well. For example, the meaning of the sentence *cats eat fishes* is related to the meaning of the words *cats*, *eat*, and *fishes*. In this paper, for the sake of simplicity, we focus exclusively on the lexicon level. This has to be considered as a first step toward the comprehension of the emergence of complex structures in languages.

2. Quantifying duality of patterning

In this section, we quantify the notion of duality of patterning as observed in real languages, defining suitable measures for the combinatoriality and compositionality (Tria, Galantucci, & Loreto, 2012).

2.1. Datasets

We considered a set of real language dictionaries. In order to encompass the largest possible lists of words, that is, including proper names and morphological derivations, we focused our attention on the list of words used by the Ubuntu spell checker. These lists are freely available on the standard repositories of Ubuntu. We chose in particular British English, French, German, Italian and Spanish.[1]

In addition to the above-mentioned lexica, we considered the list of words produced in *Human Brain Cloud* (HBC),[2] a massively multiplayer English-based word association game. HBC represents the largest available word-association database, consisting of 78,954 words and 1,474,006 associations, considerably larger than any previous word-association experiments. In previous works, a detailed discussion of the filtering procedure leading HBC to a dataset suitable for scientific purposes, as well as its detailed analysis, has been reported (Gravino, Servedio, Barrat, & Loreto, 2012). An interesting point of view is to look at the ensemble of words and associations as a complex network, where nodes are words and links are associations, and to analyze it as a word-association graph.

2.2. The phonetic transcription

Our analysis of combinatoriality and compositionality has been performed by subdividing each word of the lexicon in its phonetic components according to the International Phonetic Association[3] (IPA) coding. In order to perform an automated analysis easily, we adopted a particular encoding of the IPA alphabet, namely the Speech Assessment Methods Phonetic Alphabet[4] (SAMPA), a computer-readable phonetic script using ASCII characters. Specifically, we used eSpeak, a compact open source software speech synthesizer that translates any texts in its SAMPA encoding.

2.3. Phoneme statistics

First, we measured the distribution of word lengths in phonemes for the lexica of the five languages considered as well as for the HBC's lexicon. The normalized histogram of word lengths is reported in Fig. 1 (left). For all languages we observe fairly compatible shapes and in the same range we even see identical trends; it is worth noting that in the most important region (a zoom of which is portrayed in the inset of Fig. 1, left) all the languages show different behavior, except for the HBC and the English language. These similarities were somehow expected since HBC is practically a sample of the English language. We now define the frequency of a phoneme in a given language as the number of words containing that phoneme normalized with the total number of words in the lexicon of that language. It can also be roughly considered as the probability of appearance of a given phoneme. Fig. 1 (center) reports the frequency-rank plot of the phoneme frequencies. After rescaling the frequencies with the size of the lexicon, it is quite evident that the shape of the distributions is roughly the same, except for the number of different phonemes in each language corresponding to the maximal rank. The phonemes frequency has been measured also for the HBC's lexicon. The figure shows that HBC distribution is

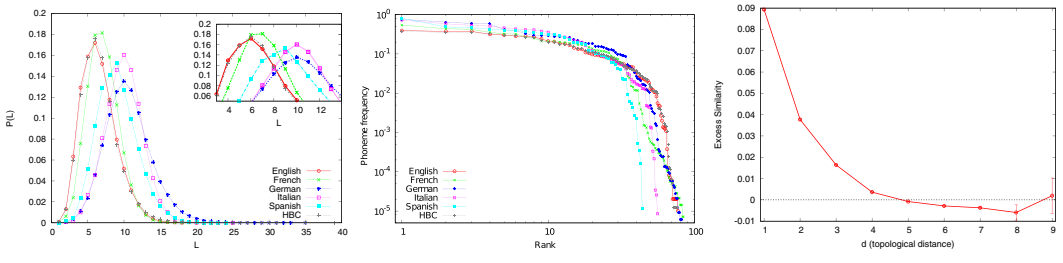

Fig. 1. *Left*: The normalized histogram of the number of phonemes composing the words of the six lexica considered. In the inset, a zoom of the region with percentages above the 5%. *Center*: The phoneme frequency-rank plot for the words of the considered lexica. On the horizontal axis, we report the rank of individual phonemes and on the vertical axis the corresponding frequency in each lexicon. *Right*: Excess Master-Mind similarity of words as a function of the average distance d of the corresponding objects on the HBC word-association graph.

quite similar to the English distribution. This fact points out again the reliability of the HBC database when used to sample the English language properties.

2.4. Entropy and combinatoriality

A complementary measure of the statistics of use of the different phonemes in the lexicon is given by the entropy. The entropy of the elementary phonemes distribution is defined as $S = -\Sigma_{f_i} p(f_i) \log(p(f_i))$ where f_i is the generic elementary form (phoneme) and $p(f_i)$ is the frequency of occurrence of f_i in the whole emerging lexicon. We adopt here a measure of combinatoriality (introduced in Tria et al. [2012], following the idea in Galantucci, Kroos, and Rhodes [2010]) to quantify the property of a communication system to combine and re-use a small number of elementary forms to produce a large number of words. To this end, we adopt a real-valued quantity ranging in the interval [0, 1] that quantifies how frequently elementary forms recur in a lexicon, according to the following expression:

$$C = \frac{\sum_i (m(f_i) - 1)}{(M - 1)F}, \tag{1}$$

where the sum runs over all the F distinct forms present in the emerged lexicon and $m(f_i)$ is the number of distinct objects whose name includes the form f_i. The term $m(f_i) - 1$ takes into account only the forms that are used to name at least two objects, that is, only the forms that are actually re-used. M is the number words in the lexicon. Table 1 reports the results for the values of the entropy and combinatoriality for all the lexica considered.

2.5. Compositionality

The aim here is that of establishing whether, in a specific lexicon, words of semantically related concepts are expressed through morphologically similar words. To this end, we need to focus on a suitable conceptual graph; that is, a network where the nodes are the concepts expressed by words and the connections are semantic relations between pairs of concepts. Here we shall consider the word-association graph produced by the HBC experiment as a proxy of the conceptual space and we shall measure the semantic similarity of two words in terms of their topological distance on the graph. In addition, we need to define a measure of morphological similarity between words. To this end we adopt the Master-Mind-like (MM) measure introduced in Tria et al. (2012). Given two words w_1 and w_2, each composed of a certain number of phonemes, the MM measure of phoneme similarity is defined as follows. After the two words have been aligned, either making the left-end or the right-end coincide, we sum 1 whenever the two words share the same form in the same position and 0.5 for the same form in a different position. We take the MM measure to be the maximum between the left-end and the right-end alignment values. This MM measure conveys the idea that meaningful forms are often included in words as suffix or prefix, and in general in a well-defined position.

To quantify compositionality, we measure (as in Tria et al., 2012) the *excess similarity* of words used to name related objects (at small topological distance in the graph) when compared to the similarity of randomly chosen words. In order to do that, we consider the average difference between the similarity between each pair of words as a function of their distance on the HBC graph and the same value computed in 10 random cases, obtained by reshuffling the associations between words and nodes in the HBC graph. Fig. 1 (right) reports the results for HBC. Ten instantiations of the null model are sufficient since the standard deviation of the distribution (represented with error bars in Fig. 1) is negligible for all meaningful distances. The excess Master-Mind similarity is larger for more semantically related words, that is, for words at small topological distance in the HBC, and it decreases monotonically with the topological distance, being significantly different from zero up to topological distances of the order of 3. We interpret these results as a clear signature of compositionality in HBC and, being HBC a good proxy for English, more generally for English. Though semantically related words are not necessarily related in forms, an excess similarity between words used to name related meanings exists in natural languages and is statistically relevant.

After this short survey of the statistics of phonemes in several lexica as well as a quantification of the duality of patterning of these lexica, we now move on a more theoretical ground, by focusing on a suitable modeling scheme aimed at explaining the emergence of the above mentioned duality of patterning.

3. Modeling the emergence of duality of patterning

We now focus on the mechanisms that could lead to the establishment of duality of patterning in a lexicon. There have been a number of previous works devoted to explaining the emergence of combinatoriality and compositionality (Beule & Bergen, 2006; Brighton, 2002; Kirby, Cornish, & Smith, 2008; Nowak, Krakauer, & Dress, 1999; Oudeyer, 2005; Smith, Brighton, & Kirby, 2003; Vogt, 2005; Zuidema & de Boer, 2009). A thorough review of the attempts presented in literature is far from the scope of the pre-

Table 1
Phoneme statistics of the different languages considered and of HBC

Word List	M	F	$S/S_{\max}$	C
British English	98,326	77	0.8228	0.079
French	139,721	84	0.7644	0.077
German	327,314	83	0.7724	0.109
Italian	116,878	56	0.8087	0.147
Spanish	86,016	44	0.8206	0.167
HBC	78,954	76	0.8336	0.080

Note. M is the number of words in the list, F is the number of different phonemes observed in the word list of each dictionary (according to the SAMPA classification), S is the entropy calculated on the normalized frequency distributions, $S_{\max} = \log(F)$ is the maximum value of the entropy reached if all phonemes were equiprobable. C is the combinatoriality defined in Eq. 1.

sent paper. Here we shall only focus on a modeling scheme, the Blending Game, recently introduced to address in a unitary framework the co-emergence of combinatoriality and compositionality (Tria et al., 2012).

3.1. The Blending Game

Here we briefly remind the definition of the Blending Game and we refer to Tria et al. (2012) for its full definition. The Blending Game focuses on the problem of how a population of N individuals can bootstrap, on cultural time-scales, a lexicon to name a set of M objects. Those M objects are not independent to each other. Rather, we hypothesize the existence of conceptual and semantic links among them, and we represent such a structure with a graph.

In this paper, we consider the set of objects to be named linked through a simple random graph (perhaps the simplest approximation one can think of), and we refer to the original paper (Tria et al., 2012) for an extensive analysis of different kinds of graph. This choice also allows modeling the effect of differently shaped conceptual spaces and of conceptual spaces that may differ from individual to individual.

We consider an initially *blank slate* population of N individuals committed to bootstrapping a lexicon by using an open-ended set of forms in a large conceptual space composed by M objects to be named. In the previous section, we focused on phonemes, that is, elementary forms, while in the following we shall refer to generic forms, intended as meaningless units. For simplicity, we consider here a population of agents with a homogeneous structure, where each agent (individual) has the same probability of interacting with everyone else.

As we said, we consider the M objects to be named as related through a graph structure encoding their conceptual/semantic relations. In this representation, we imagine that two objects linked in the graph are conceptually related. For sake of brevity we refer to this graph as the conceptual network. All the results we present below are obtained with a conceptual network modeled as an Erdös–Rényi (ER) random graph (Erdös & Rényi, 1959). ER graphs are constructed as follows. One starts with a set of M unconnected nodes. Then for each node one draws a link between this node and all the other $M - 1$ nodes, each with probability p_{link}. The outcome is a graph with and average number of connections per node of the order of $M \cdot p_{\text{link}}$. Here p_{link} is controlling how dilute is the graph, that is, small values of p_{link} corresponding to graphs with less links.

Starting from scratch and without any central coordination, the individuals perform pairwise language games (Baronchelli, Felici, Caglioti, Loreto, & Steels, 2006; Puglisi, Baronchelli, & Loreto, 2008; Steels, 1995), where a randomly selected pair of Speaker (S) and Hearer (H) a communication act aimed at naming objects of their conceptual space. In order to study the emergence of duality of patterning, we consider the possibility for any word to be composed either of a single form or of an ordered linear concatenation of forms.

The Blending Game features two main ingredients: (a) noise in comprehension and (b) a blending repair strategy. Noise is such that H correctly understands each form uttered

by S with a time dependent probability $P_t(f_i) = (1 - \exp(-\frac{n_t(f_i)}{\tau}))$, where $n_t(f_i)$ is the number of times H has heard the form f_i up to the current time t, and τ is a characteristic memory scale. This is an essential ingredient responsible of keeping the number of different forms shared by the whole population limited and low, without any a priori constraint on it. The blending repair strategy exploits the structure of the world to create new words. In fact, if the communication fails, the S composes a new word by reusing parts of already proposed words, preferably associated with object nearest neighbors of the chosen topic. This Blending Repair strategy can be thought of as a tinkering process (Zuidema & de Boer, 2009) through which language users recombine already existing materials instead of inventing them from scratch. We again refer to Yule (1944) for a comprehensive description of the game.

The dynamics of the Blending Game are such that the communicative success rate starts from zero and progressively increases monotonously, leading eventually to a fully successful shared communication. The emerging asymptotic lexicon does not present homonymy and is such that each object is associated with a word composed by a certain number of forms.

3.2. Properties of the emerged lexicon

Owing to the blending procedure, the lexicon shared by the population contains words that are composed by several forms. A typical word in the lexicon is, for instance, $f_{23}f_{18}f_0f_0$, composed by the elementary forms f_{23}, f_{18} and f_0. We here consider the length of a word as the number of forms composing it. In Fig. 2 (left), the word length distribution in the emerging lexicon is reported. We note that the observed limitation on the word length is an outcome of the negotiation dynamics, emerging without any explicit constraints on it. The distribution features a typical shape that has been observed in human languages (Sigurd, Eeg-Olofsson, & Van Weijer, 2004), well fitted by the function $f(x) = ax^bc^x$, which corresponds to the Gamma distribution when the parameters are suitably renamed (Sigurd et al., 2004). Fig. 2 (left) shows the word length distribution for different values of the rescaled memory parameter $\tau_M = \tau/M$ and of the rescaled graph connectivity p_M (see the caption of Fig. 2 for the definition). The average word length in the emerging lexicon is thus very stable when changing the parameters of the model, remaining finite and small, and comparable with the length of words in human languages (compare with Fig. 1).

3.3. Frequency-rank distribution of elementary forms

As a deeper investigation of the properties of the emerged lexicon, we consider the frequency-rank distribution of the different forms composing the words (Fig. 2, center and right). As in the case of the word length distribution, the frequency-rank distribution for forms does not depend on p_{link} (see Fig. 2, right). However, we note in this case a clear dependence on the memory parameter τ_M. In particular, the higher τ (for M fixed); that is, the lower the fidelity in communication, the smaller the number of distinct forms on which

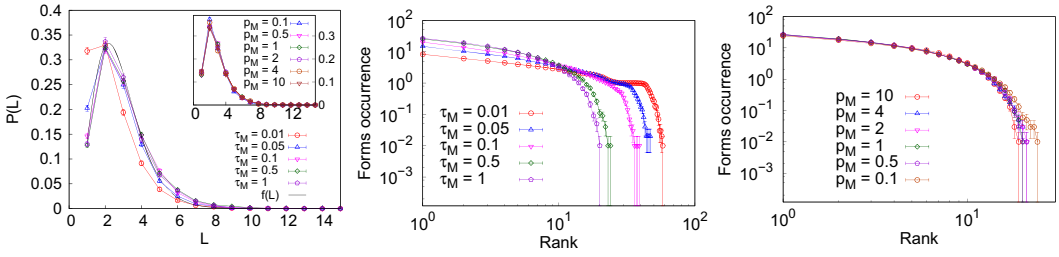

Fig. 2. *Left*: Distribution of word length L for different values of τ and p_{link} (the linking probability of ER graphs). In order to compare graphs corresponding to different sizes M, of the conceptual graph, we define the following normalized quantities. A normalized link probability for the ER random graph as $p_M = \frac{p_{\text{link}}}{p^*}$, where $p^* = \frac{\log M}{M}$ is the threshold above which the whole graph in connected with probability one in the infinite size limit ($M \to +\infty$). Similarly, a normalized time scale parameter is defined as $\tau_M = \frac{\tau}{M}$. The distribution of word length $P(L)$ in the figure is obtained by fixing $p_M = 0.5$. $f(L) = aL^b c^L$ is the observed empirical function (Sigurd et al., 2004) (solid black line, with fitting parameters $a = 0.74$, $b = 3.7$, $c = 0.18$). In the top inset we show the same distribution fixing $\tau_M = 1$ for different values of p_M. Note that here the curves overlap, indicating that the word length distribution does not depend on the object graph connectivity. The number of agents and the number of objects in the environment are fixed respectively to $N = 10$ and $M = 40$. *Center*: frequency-rank distribution for elementary forms is shown for different values of the parameter τ_M keeping fixed $p_M = 0.5$, $N = 10$ and $M = 40$. *Right*: the same distribution fixing $\tau_M = 1$ and for different p_M, showing that the distribution of elementary forms does not depend on the object graph connectivity. All the results shown are averaged over 100 different realizations of the process on the same conceptual graph. The figure has been adapted from Tria et al. (2012).

the agents eventually find agreement. Since the invention rate of new forms does not depend on τ, the effect of a high τ is that of strengthening the selection process, reducing in this way the number of forms that get fixed in the population. It is worth noticing that for large values of τ, the frequency-rank distribution we observe is remarkably similar to the corresponding distribution observed in human languages (Tambovtsev & Martindale, 2007) for which a Yule-like distribution (Yule, 1944) has been hypothesized.

3.4. Combinatoriality

We now compute the combinatoriality defined in Eq. 1 to quantify how frequently forms recur in the emerged lexicon. We report the results in Fig. 3 (left) as a function of τ_M and for different values of p_M. Again, a negligible dependence on p_{link} is found, while, as in the case of the frequency-rank distribution, a clear dependence on τ_M is found, the maximal combinatoriality occurring for high values of τ_M. This can be understood if one thinks that for a perfect level of understanding there is no selective pressure acting on the different forms and many distinct forms are eventually fixed in the lexicon with a small reuse rate, that is, little combinatoriality. Summarizing, when the effort for understanding new forms is sufficiently high, one finds, at the lexical level, features similar to the ones observed in real languages, such as the word length distribution and the number and the frequency of use of the different forms. In this perspective, combinatoriality emerges as a

workaround to overcome the problem of noisy communication. Interestingly the values predicted by the Blending Game for the combinatoriality are not far from those measured in real languages (see Table 1).

3.5. Compositionality

Let us now turn to the compositional aspects of the lexicon, again aiming at establishing whether, in the emerged lexicon, words for semantically related concept are expressed through morphologically similar words. We refer again to the definition of *excess similarity* as introduced above. In Fig. 3 (right), we report the excess similarity for a fixed value of τ_M and several values of p_M as a function of the topological distance d on the conceptual graph. Compositionality is evident in the figure: the more closely related the words, the higher the excess similarity. It is also remarkable the qualitative similarity with the corresponding measure performed in real languages is also remarkable (see Fig. 1, right). At odds with the above studied properties of the lexicon, the excess similarity only weakly depends on τ_M, while it strongly depends on p_M. This indicates that a percolation of the organization of the world into the lexicon is possible when the world has a non-trivial semantic structure, that is, in our case when p_{link} is different from zero and from one. In the former case no relation between objects exists, while in the latter case all the objects are equally related (all are at distance one in the graph). Sparse graphs are more prone to induce compositionality.

4. Discussion and perspectives

In this paper, we have investigated duality of patterning at the lexicon level. In particular, we have quantified the notions of combinatoriality and compositionality as observed in real languages as well as in a large-scale dataset produced in the framework of a web-based word association experiment (Gravino et al., 2012). We have paralleled this empirical analysis with a modeling scheme, the Blending Game, whose aim is that of identifying the main determinants for the emergence of duality of patterning in language. We analyzed the main properties of the lexicon emerged from the Blending Game as a function of the two parameters of the model, the graph connectivity p_{link} and the memory scale τ. We found that properties of the emerging lexicon related to the combinatoriality, namely the words length distribution, the frequency of use of the different forms, and a measure for the combinatoriality itself, reflect both qualitatively and quantitatively the corresponding properties as measured in human languages, provided that the noise parameter τ is sufficiently high and that a sufficiently high effort is required in order to understand and learn brand new forms. Conversely, the compositional properties of the lexicon are related to the parameter p_{link}, measure of the level of structure of the conceptual graph. For intermediate and low values of p_{link}, semantic relations between objects are more differentiated with respect to the situation of a more dense graph, in which every object is related to anyone else, and compositionality is enhanced. In summary, while

the graph connectivity strongly affects the compositionality of the lexicon, noise in communication strongly affects the combinatoriality of the lexicon.

These results are important because they demonstrate for the first time that the two sides of duality of patterning can emerge simultaneously as a consequence of a purely cultural dynamics in a simulated environment that contains meaningful relations. Two main ingredients seem to be crucial to this mechanism: the existence of noise in comprehension and a blending repair strategy that combines and re-uses forms taken from other object's names. This last point is reminiscent of the tinkering mechanism proposed by Jacob (1977), and it could bridge a link to investigate how recent advances in evolutionary developmental biology (Carroll, 2005) (evo-devo in short) could apply to the emergence and evolution of language features, addressing important questions as: What are the mechanisms to introduce innovations (recombination (Arthur, 2008; Schumpeter, 1934), tinkering (Jacob, 1977), exaptation (Gould & Vrba, 1982), adjacent possible (Tria, Loreto, Servedio, & Strogatz, 2015), and how new features can get successful and stabilize at the population level (for similar modeling scheme incorporating truly evolving populations see Colaiori et al., 2015). From this perspective, we believe that the modeling scheme illustrated here could be an important framework to export evo-devo concepts in studies about language evolution.

Many directions are open for future investigations. First of all, one could elucidate the emergence of duality of patterning at the syntax level beyond that of the lexicon. In addition, many different modifications of our modeling scheme are possible. A very interesting one consists in relaxing the assumptions that the conceptual space of all the individuals is identical and modeled as a static graph, imaging instead that the conceptual space of each individual gets continuously reshaped by the interactions among the users. In this way one would realize a true co-evolution of the conceptual spaces of the individ-

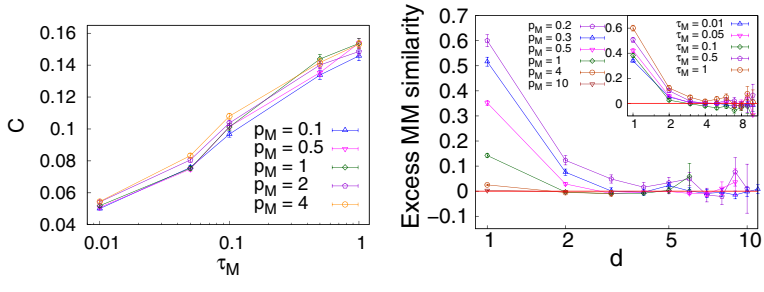

Fig. 3. *Left*: Combinatoriality C for different values of p_M as a function of τ_M. The number of agents and the number of objects in the environment are fixed respectively to $N = 10$ and $M = 40$. *Right*: Excess Master-Mind-like similarity of words as a function of the distance d of the corresponding objects on the graph. A decrease in the excess similarity as a function of the topological distance d is the signature of the emergence of compositionality; in particular, compositionality implies higher similarity among words which are closer in the semantic space. The topological distance on the object graph is our proxy for the semantic relatedness. Results are reported for $N = 10$ and $M = 100$ and for different values of the objects graph connectivity p_M, keeping fixed $\tau_M = 1$ (main figure) and for different values of τ_M keeping fixed $p_M = 0.2$ (inset). The figure has been adapted from Tria et al. (2012).

uals and of their inventories of associations between objects and words. Finally, it is worth mentioning how recent advances in information and communication technologies allow the realization of focused experiments also in the framework of the emergence of linguistic structures, and a general trend is emerging for the adoption of web-games (see, for instance, the recently introduced Experimental Tribe platform http://www.xtribe.eu; Caminiti et al., 2013) as a very interesting laboratory to run experiments in the social-sciences and whenever the contribution of human beings is crucially required for research purposes. This is opening tremendous opportunities to monitor the emergence of specific linguistic features and their co-evolution with the structure of conceptual spaces.

Acknowledgments

It is our pleasure to warmly thank Bruno Galantucci for very interesting scientific discussions. The authors acknowledge support from the KREYON project funded by the Templeton Foundation under contract N. 51663 and the EuroUnderstanding Collaborative Research Projects DRUST funded by the European Science Foundation.

Notes

1. Downlodable at http://www.debian.org/distrib/packages
2. Dataset courtesy of Kyle Gabler, http://kylegabler.com/
3. http://www.langsci.ucl.ac.uk/ipa
4. http://www.phon.ucl.ac.uk/home/sampa/

References

Arthur, W. B. (2008). *The nature of technology*. New York: The Free Press.

Baronchelli, A., Felici, M., Caglioti, E., Loreto, V., & Steels, L. (2006). Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics*, *6*, P06014.

Beule, J. D., & Bergen, B. K. (2006). On the emergence of compositionality. In A. Cangelosi (Ed.), *Proceedings of the 6th international conference on the evolution of language* (pp. 35–42). Rome: World Scientific.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, *8*(1), 25–54.

Caminiti, S., Cicali, C., Gravino, P., Loreto, V., Servedio, V. D. P., Sirbu, A., & Tria, F. (2013). Xtribe: A web-based social computation platform. *Third International Conference on Cloud and Green Computing (CGC)*, 397–403.

Carroll, S. B. (2005). *Endless forms most beautiful: The new science of evo devo and the making of the animal kingdom*. New York: WW Norton & Company.

Colaiori, F., Castellano, C., Cuskley, C. F., Loreto, V., Pugliese, M., & Tria, F. (2015). General threestate model with biased population replacement: Analytical solution and application to language dynamics. *Physical Review E*, *91*, 012808.

Erdös, P., & Rényi, A. (1959). On random graphs I. *Publications Mathematics Debrecen*, *6*, 290.

Galantucci, B., Kroos, C., & Rhodes, T. (2010). The effects of rapidity of fading on communication systems. *Interaction Studies*, *11*(1), 100–111.

Gould, S. J., & Vrba, E. S. (1982). Exaptation: A missing term in the science of form. *Annales de l'I.H.P.*, *8*(1), 4–15.

Gravino, P., Servedio, V. D. P., Barrat, A., & Loreto, V. (2012). Complex structures and semantics in free word association. *Advances in Complex Systems*, *15*, 1250054.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*, 88–96.

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*, 1161–1166.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(31), 10681–10686.

Nowak, M. A., Krakauer, D., & Dress, A. (1999). An error limit for the evolution of language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, *266*(1433), 2131–2136.

Oudeyer, P.-Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*, *233*(3), 435–449.

Puglisi, A., Baronchelli, A., & Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(23), 7936.

Schumpeter, J. (1934). *The theory of economic development*. Cambridge, MA: Harvard University Press.

Sigurd, B., Eeg-Olofsson, M., & Van Weijer, J. (2004). Word length, sentence length and frequency—zipf revisited. *Studia Linguistica*, *58*(1), 37–52.

Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: the cultural emergence of compositional structure. *Advances in Complex Systems*, *6*(4), 537–558.

Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial Life*, *2*(3), 319–332.

Tambovtsev, Y., & Martindale, C. (2007). Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics*, *4*(2), 1–11.

Tria, F., Galantucci, B., & Loreto, V. (2012). Naming a structured world: A cultural route to duality of patterning. *PLoS ONE*, *7*, e37744.

Tria, F., Loreto, V., Servedio, V. D. P., & Strogatz, S. H. (2015). The dynamics of correlated novelties, Nature Scientific Reports No. 4:5890.

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, *167*(1-2), 206–242.

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge, England: Cambridge University Press.

Zuidema, W., & de Boer, B. (2009). The evolution of combinatorial phonology. *Journal of Phonetics*, *37*(2), 125–144.