ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Check for updates

Cite this article: Corominas-Murtra B, Sànchez Fibla M, Valverde S, Solé R. 2018 Chromatic transitions in the emergence of syntax networks. *R. Soc. open sci.* **5**: 181286. http://dx.doi.org/10.1098/rsos.181286

Received: 3 August 2018
Accepted: 29 October 2018

Subject Category:
Psychology and cognitive neuroscience

Subject Areas:
biophysics/psychology/computational biology

Keywords:
complex networks, graph colouring, modularity, syntax

Authors for correspondence:
Bernat Corominas-Murtra
e-mail: bernat.corominas-murtra@ist.ac.at
Ricard Solé
e-mail: ricard.sole@upf.edu

THE ROYAL SOCIETY
PUBLISHING

Chromatic transitions in the emergence of syntax networks

Bernat Corominas-Murtra[1], Martí Sànchez Fibla[2], Sergi Valverde[3,4] and Ricard Solé[3,4,5]

[1]Institute of Science and Technology Austria, Am Campus 1, 3400, Klosterneuburg, Austria
[2]Technology Department, Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona, Spain
[3]Complex Systems Lab, ICREA-Universitat Pompeu Fabra, Dr Aiguader 88, 08003 Barcelona, Spain
[4]Evolution of Technology Lab, Institut de Biologia Evolutiva (CSIC-UPF), Passeig Maritim de la Barceloneta, 37-49, 08003 Barcelona, Spain
[5]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

BC-M, 0000-0001-9806-5643; RS, 0000-0001-6974-1008

The emergence of syntax during childhood is a remarkable example of how complex correlations unfold in nonlinear ways through development. In particular, rapid transitions seem to occur as children reach the age of two, which seems to separate a two-word, tree-like network of syntactic relations among words from the scale-free graphs associated with the adult, complex grammar. Here, we explore the evolution of syntax networks through language acquisition using the *chromatic number*, which captures the transition and provides a natural link to standard theories on syntactic structures. The data analysis is compared to a null model of network growth dynamics which is shown to display non-trivial and sensible differences. At a more general level, we observe that the chromatic classes define independent regions of the graph, and thus, can be interpreted as the footprints of incompatibility relations, somewhat as opposed to modularity considerations.

1. Introduction

The origins of human language have been a matter of intense debate. Language is a milestone in our evolution as a dominant species and is likely to pervade the emergence of cooperation and symbolic reasoning [1–4]. Maybe the most defining and defeating trait is its virtually infinite generative potential: words and sentences can be constructed in recursive ways to generate nested structures of arbitrary length [3,5]. Such structures are the product of a set of rules defining syntax, which are extracted by human brains through language acquisition during childhood after a small sample of the whole combinatorial universe of sentences has been learned. And yet, in spite of its complexity, syntax is accurately acquired by children, who master their mother tongue in a few years of learning. Indeed, around the age of two, linguistic structures produced by children display a qualitative shift on their complexity, indicating a deep change on the rules underlying them [6–8]. This sudden increase of grammar complexity is known as the *syntactic spurt*, and defines the edge between the *two words* stage, where only isolated words or combinations of two words occur, to a stage where the grammar rules governing this syntax are close to the one we can find in adult speech—although the cognitive maturation of kids makes the semantic content or the pronunciation different from the adult one. How can we explain or interpret such nonlinear pattern?

Statistical physicists have approached the problem of language evolution showing, for example, that non-trivial patterns are shared between language inventories—collections of words—and some genetic and ecological neutral models [9] (see [10] and references therein). However, most of these models do not make any assumption about the role played by actual interactions among words, or, more generally, linguistic units, which largely define the nature of linguistic structures. In this context, a promising approach to its structure and evolution involves considering language in terms of networks of interconnected units instead of unstructured collections of elements—e.g. words or syllables [11–15]. In this context, syntactic networks, in which nodes are words and links the projection of actual syntactic relations, have been shown to be an interesting abstraction to grasp general patterns of language production [7,8,16]. Specially valuable has been the quantitative data obtained from syntax networks obtained along the process of syntax acquisition, for they provided solid and quantitative evidence of sudden qualitative shifts in the cognitive machinery involved in the process, present also in other linguistic domains [7,8,16–19].

At the fundamental level, syntax can be understood as a set of symbols associated under a universe of potential combinations somewhat similar to chemistry. Atoms and words would then be linked through compatibility relations defining what can be combined and what is forbidden. The power of this picture is supported by the use of linguistic methods in the systematic characterization of chemical structures [20]. Chemical structure diagrams can thus be seen as some sort of language, with chemical species and bonds as key ingredients. In a more abstract fashion, we can say that general rules of combining elements within a given set of interacting pieces with well-defined functional meaning is at work in both language and chemistry. Following the chemical analogy, where abstract classes of 'nodes' can be defined, we will take advantage of graph colourability theory as a general framework to detect transitions based on qualitative changes of compatibilities. Specifically, we suggest that a new combinatorial approach grounded on graph colouring may enable a better understanding of the evolution of networks having internal relations of compatibility—e.g. some kind of syntactic rules. In this context, we propose the chromatic number—and complementary measures—of the graph [21–23] as an indicator of network complexity. The chromatic number is defined as the minimal number of *colours* needed to *paint* all nodes of the graph in a way that no adjacent nodes have the same colour [22]. In other words, classes of nodes would be defined precisely by the fact that there are no connections among them, *a measure conceptually opposite to graph modularity*. The q-colouring problem, i.e. to know whether a graph can be coloured with q different colours, is one of the most important NP-complete problems. From the statistical physics point of view, an analogous problem is defined within the context of the *Potts* model [24]. Transitions in the evolution of the chromatic number, which is the main objective of this work, have been widely studied in abstract models of random graphs [23,25–27].

The chromatic number may convey structural information among the classes of relations the graph is showing within the system it aims to abstract. This is exactly the point by which graph colouring is relevant for syntactic phenomena. We will work with a graph of aggregated syntactic relations using the paradigm of *dependence grammar* [28], which provides a natural framework to extract graphs from syntactic relations [7,8]. The aim of using the chromatic number comes from the intuition that syntactic relations do not glue elements *for free* but display consistent rules of compatibility/incompatibility among lexical elements. This may seem obvious at the level of the sentence analysis, but to extrapolate how these combinatorial rules among different classes of elements work at the global system level is a hard task, and even harder, if we want to do it quantitatively. For example, to grasp the relevance of chromatic number, one must perform parallel measurements on the network using indicators taking into account potential deviations—which will be the footprints of the non-trivial compatibility relations. The interplay between the evolution of the chromatic number and its deviations from a null model made of free associations will be the target of our paper. As we shall see, non-trivial transitions between different, increasing chromatic numbers, along with interesting deviations from a null model of syntax-free sentence generation are identified. This is, to the best of our knowledge, the first time that such transitions have been reported in a real system.

2. Graphs and colouring: basics

We will work over undirected graphs. An undirected graph $\mathcal{G}(V, E)$—hereafter, $\mathcal{G}$—is composed by the set of $V = \{v_1, \ldots, v_n\}$ *nodes* and a set $E = \{e_j \| 1 \leq j \leq m\} \subseteq V \times V$ of edges. Each (unordered) pair $e_j = \{v_i, v_k\}$ depicts a link between nodes v_i and v_j. The number of links $k(v_i)$ attaching node v_i is the degree of the node and $\langle k \rangle$ is the *average degree* of the graph $\mathcal{G}$. The *degree distribution* $P(k)$ accounts for the probability to select a node at random having degree k. The identity card of a graph is the so-called *adjacency matrix*, $\mathbf{a}(\mathcal{G})$, which is defined as follows:

$$a_{ij} = \begin{cases} 1, & \text{iff } (\exists e_k \in E) : (e_k = \{v_i, v_j\}) \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

We observe that the adjacency matrix of undirected graphs is symmetrical, i.e. $a_{ij} = a_{ji}$.

We can map the chromatic problem into the antiferromagnetic q-dimensional Potts model at $T = 0$ [24]. This model is a generalization of the classical Ising model for lattices: at every node of this lattice we place a particle having a spin which energetically constrains the state of its neighbours. Traditionally, spins can have only two states, namely $|\uparrow\rangle$ and $|\downarrow\rangle$. In the Potts model, compatibility relations take into account an arbitrary number $q > 2$ of different states. Let us consider a partition of nodes V containing q different classes, namely, $G_q(V) = \{g_1, \ldots, g_q\}$ of V, i.e.

$$\bigcap G_q = \emptyset \quad \text{and} \quad \bigcup G_q = V, \tag{2.2}$$

The *state* σ_i of node v_i indicates the class of $G_q(V)$ to which the node belongs, i.e. $\sigma_i \in g_j$. Let $\mathcal{F}_q(V)$ be the ensemble of all partitions of V containing q different classes. Every element in $\mathcal{F}_q(V)$ has the following energy penalty[1]:

$$\mathcal{H}(G_q) = J \sum_{i<j} a_{ij} \delta(\sigma_i, \sigma_j), \tag{2.3}$$

where $J = 1$ is the *coupling constant* and δ is the Kronecker symbol

$$\delta(\sigma_i, \sigma_j) = \begin{cases} 1, \text{iff } i = j \\ 0, \text{otherwise.} \end{cases} \tag{2.4}$$

Intuitively, the higher the presence of pairs of connected nodes belonging to the same state, the higher will be the energy of the global state of the graph. Given a fixed q, the configurations displaying minimal energy may have an amount of non-solvable situations, leading to the unavoidable presence of connected nodes at the same state. This phenomenon is called *frustration*, and for these configurations, the ground state of the Hamiltonian defined in (2.3) displays positive energy. If there is no frustration, i.e. $\exists G_q \in \mathcal{F}_q(V)$, we can find a partition that satisfies

$$\mathcal{H}(G_q) = 0, \tag{2.5}$$

and we say that the graph is q-colourable, being the q different *colours* the q different classes or members of G_q. When the graph is q-colourable, there is at least one partition $G_q \in \mathcal{F}_q(V)$ such that, if $v_i, v_j \in V$ belong to the same *class* or *colour* of the partition, namely $g_l \in G_q$. We deduce that

$$(v_i, v_j \in g_l) \;\Rightarrow\; a_{ij} = 0. \tag{2.6}$$

Relation (2.6) maps colour classes onto disjoint sets of graph elements (adjacent nodes have a different colour). Now, the colouring problem consists in finding the minimal number of classes (or colours) required to properly *paint* the graph. This is the so-called *chromatic number* of the graph $\mathcal{G}$

$$\chi(\mathcal{G}) = \min\{q : (\exists G_q \in \mathcal{F}_q(V)) : \mathcal{H}(G_q) = 0\}. \tag{2.7}$$

Now suppose network partition(s) $G_q^* \in \mathcal{F}_q(V)$ having minimal energy, see equation (2.3), given a number of colours q

$$G_q^* = \min_{G_q \in \mathcal{F}_q(V)} \{\mathcal{H}(G_q)\}. \tag{2.8}$$

[1] In our approach, the energy units of this Hamiltionian are arbitrary.

In general, the process of search for the chromatic number yields a decreasing sequence of energies ending at $\mathcal{H}(G^*_{\chi(\mathcal{G})}) = 0$

$$\mathcal{H}(G^*_1) \leq \cdots \leq \mathcal{H}(G^*_{\chi(\mathcal{G})}) = 0, \tag{2.9}$$

In order to assess the statistical significance of chromatic numbers, we define the *relative energy* of any q-colouring as follows:

$$f_q(G^*_q) = \frac{\mathcal{H}(G^*_q)}{|E|}, \tag{2.10}$$

where $|E|$ is the number of edges in the graph G. This quantity $0 \leq f_q(G^*_q) \leq 1$ corresponds to the minimal (relative) number of frustrated links or *violations* (i.e. when adjacent nodes have the same colour).

Despite the high complexity of this problem—computing the chromatic number in an arbitrary graph is a *NP*-hard problem—several bounds can be defined. A lower bound can be defined from the so-called *clique number*. A *clique* is a subgraph in which every node is connected to all other nodes in the subgraph. The *clique number* $\omega(\mathcal{G})$ is the size of the largest clique in the graph, which is a natural lower bound for $\chi(\mathcal{G})$ [22]

$$\omega(\mathcal{G}) \leq \chi(\mathcal{G}). \tag{2.11}$$

Alternatively, an upper bound on $\chi(\mathcal{G})$ can be defined by looking at the K-core structure of $\mathcal{G}$. The $K(\mathcal{G})$ core is the largest subgraph whose nodes display degree higher or equal to K. Now, let $K^*(\mathcal{G})$ be the K-core with largest connectivity that can be found in $\mathcal{G}$

$$K^* = \max\{K : K(\mathcal{G}) \neq \emptyset\}. \tag{2.12}$$

Then, it can be shown that K^* sets an upper bound to the chromatic number [22]

$$\chi(\mathcal{G}) \leq K^* + 1. \tag{2.13}$$

Finally, let us mention that, for some families of random graphs the chromatic number has an asymptotic behaviour depending on the average connectivity [23], $\chi(\mathcal{G}) \sim \langle k \rangle / \log\langle k \rangle$. However, the above relationship does not hold for scale-free networks with exponent $2 < \gamma < 3$. These heterogenous networks cannot have a stable value of the chromatic number because their clique number (2.11) diverges with the graph size, even at constant $\langle k \rangle$ [29].

3. The evolution of χ along syntax acquisition

Here, we study the evolution of the chromatic number through language development as captured by syntax graphs. We compare the chromatic number with the lower and upper bounds provided by the clique number and the maximal K-core, respectively. We assess the relevance of computed chromatic numbers with the corresponding minimal energy. The combination of these two measurements enable us to interpret the nature of the chromatic number. Specifically, we can check whether changes in this number reflect a global pattern or instead some anomalous behaviour of a small, localized subgraph. Finally, we provide further validation of our analysis by comparing chromatic numbers in empirical and synthetic networks obtained through a random sentence generator.

3.1. Building the networks of early syntax

Through the process, networks built upon the aggregation of syntactic structures from child's productions grow and change in a smooth fashion until a rapid transition occurs [7,8,30,31] (see also [13]). We reconstruct syntactic networks by projecting the raw constituent structure, i.e. phrase structure of children's utterances, into linear relations among lexical items, in what is known as *dependency grammar* analysis [28,32]. Then, we aggregate all these productions in a single graph where nodes are lexical items and links represent syntactic relations between them [7,30,31]. We emphasize that these networks have been built *by hand*, in the sense that no automatic procedure has been at work. The reason stems from the fact that early child language is far from normative, but, still, structures can be identified. Therefore, each link is discussed after checking its suitability according to specific linguistic criteria developed for this analysis (see [7,30,31] and references therein). These networks provide a unique window into the patterns of change occurring in the language acquisition process.

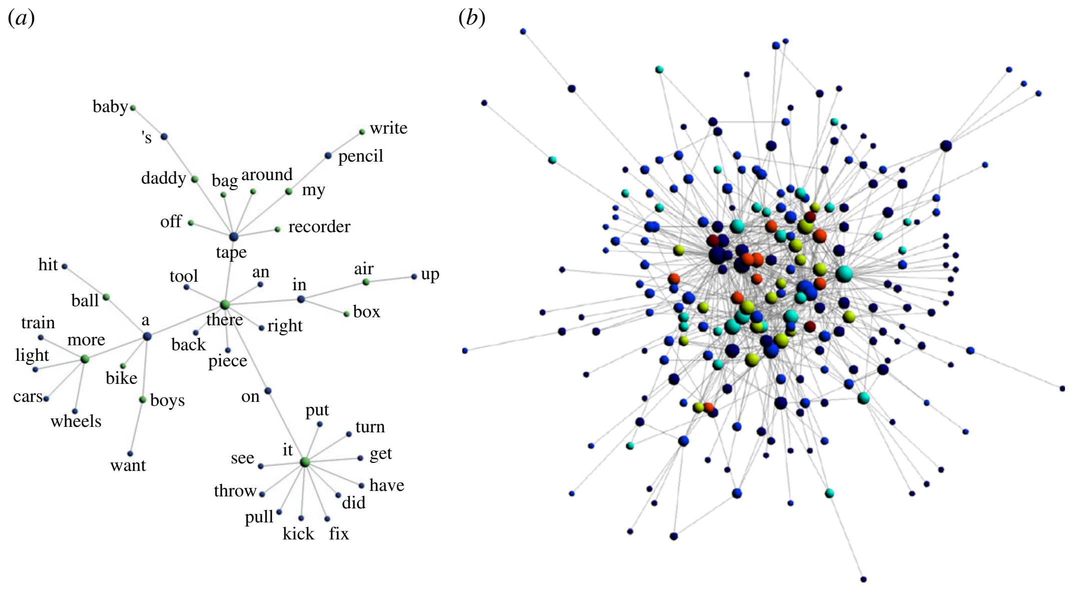

Figure 1. Optimal colourings of syntactic networks before and after the syntactic spurt. (*a*) A syntactic network before the transition (3rd corpus) is largely bipartite (this network accepts a 2-colouring). (*b*) Post-transition network (7th corpus) is remarkably more complex, which corresponds to high chromatic number $\chi(\mathcal{G}_7) = 6$. All networks coming from Peter dataset. Time spent between these two corpora is about two and a half months—see text.

The two cases studied here are obtained from the CHILDES database [33,34] which includes conversations between children and parents. Specifically, we choose Peter and Carl's corpora, whose structure has been accurately extracted and curated. For both Peter and Carl's corpora, we choose 11 different recorded conversations distributed in approximately uniform time intervals ranging from the age of approximately 20 months to the age of approximately 28 months. The chosen interval corresponds to the period in which the syntactic spurt takes place. From every recorded conversation, we extract the syntactic network of child's utterances obtaining a sequence of 11 syntactic graphs corresponding to the sequence of Peter's conversations $\mathcal{G}_{P1}, \ldots, \mathcal{G}_{P11}$ and Carl's conversations $\mathcal{G}_{C1}, \ldots, \mathcal{G}_{C11}$.

3.2. Chromatic transition from bipartite to multicoloured networks

From our graph collection (see §3.1), we obtain two sequences of chromatic numbers $s_P(\chi)$ and $s_C(\chi)$ corresponding to the evolution of the chromatic number in Peter and Carl datasets, respectively:

$$s_P(\chi) = \chi(\mathcal{G}_{P1}), \ldots, \chi(\mathcal{G}_{P11})$$
$$s_C(\chi) = \chi(\mathcal{G}_{C1}), \ldots, \chi(\mathcal{G}_{C11}).$$

The above sequences display similar patterns with some interesting differences (figures 2 and 4). For example, the middle stages of both datasets show an increase in the chromatic number. At the stage when the syntactic spurt takes place, Peter's dataset s_P displays a sharp transition from a nearly constant, low chromatic number ($\chi = 2$ up to just before month 23) to a high chromatic number (up to $\chi = 6$, month 25) which is fully consistent with the emergence of complex syntax. The first three networks in s_P accept 2-colourings, i.e., they are bipartite, see figure 1.

The grammar at this stage mainly generates pairs of complementary words, like:

$$\langle \text{verb}, \ \text{noun} \rangle \text{ or}$$
$$\langle \text{adjective}, \ \text{noun} \rangle.$$

Typical productions of this stage are, for example, 'car red' or '*horsie* run'. This pre-transition pattern, also called 2-word stage, corresponds to a highly restrictive grammar, e.g. syntactic structures like ⟨verb, verb⟩ do not exist. Instead, relations between lexical items are strongly constrained by their semantic content. On the other hand, Carl's sequence s_C shows $\chi \geq 3$ from the very beginning—i.e. these networks are not bipartite. A detailed inspection of Carl's productions at this stage shows the presence of functional particles from the very beginning. Functional particles are those lexical items whose role is essentially grammatical, and whose appearance must be accompanied by another, strongly semantic word, like a

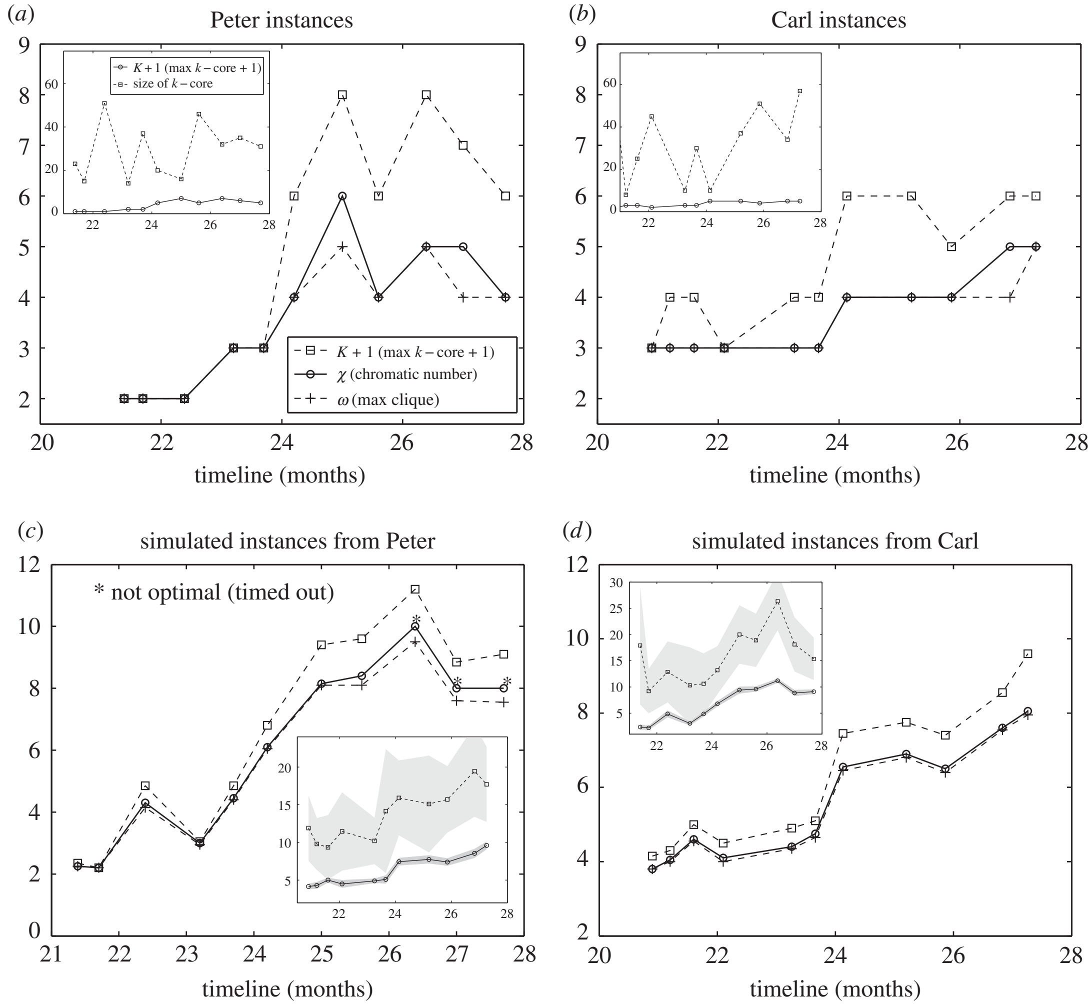

Figure 2. Evolution of the chromatic number χ (solid line), maximum clique ω (dashed line with crosses) and the maximum core $K^* + 1$ (dashed line with squares) in Peter's corpus (*a*) and Carl's corpus (*b*). Evolution of the same measures over an ensemble of $n = 20$ networks obtained after running the null model fed with Peter's corpus data (*c*) and Carl's corpus (*d*). Insets: comparison between the time evolution of the deepest *K*-core, K^*-core, size (dashed line)—i.e. number of nodes within this subgraph—and the K^* corresponding to the deepest *K*-core (solid line) for each set of networks. Shaded grey areas correspond to standard deviation in the case of the simulated instances.

noun or a verb. We consider as *functional particles* the set of lexical items composed by articles—like *a* or *the*, prepositions—like *at* or *with*, auxiliary verbs—like *do* or *will*, when they accompany another verb. The presence of functional particles from the very beginning in Carl's corpus suggests that, in general, high chromatic numbers relate to high grammar flexibility, this flexibility being provided by the *hinge* role that these particles have in the global functioning of grammar.

Still, the behaviour of the chromatic number of a graph $\chi(\mathcal{G})$ can be quite sensitive to the anomalous behaviour of small subgraphs. For example, the transition of $\chi(\mathcal{G}_2) = 2$ to $\chi(\mathcal{G}_3) = 3$, when Peter is about 23 months old, is due to a single triangle in a (largely) bipartite network (figure 2*a*,*c*). A combination of measurements enables us to assess whether the chromatic number represents the behaviour of a small number of nodes or is the natural outcome of global network features. Our choice is to compare $\chi(\mathcal{G})$ with the lower bound given by the clique number (2.11) and the upper bound provided by the maximal *K*-core connectivity (2.12). Therefore, each sequence $s_P(\chi)$, $s_C(\chi)$ will be accompanied by two sequences, namely Ω, κ

$$\Omega_{P,C} = \omega(\mathcal{G}_{P1,C1}), \ldots, \omega(\mathcal{G}_{P11,C11})$$
$$\kappa_{P,C} = K^*(\mathcal{G}_{P1,C1}), \ldots, K^*(\mathcal{G}_{P11,C11}).$$

For example, figure 2*a*,*b* shows a clear increasing trend both for maximum clique and maximum *K*-core. This, combined with the sequence of energy values given in table 1, indicates that the final chromatic

Table 1. Relative energy values of q-colourings in the Peter (top) and Carl (bottom) datasets. Relative energies reveal the fraction of frustrated links in the optimal colouring using q different colours.

	$\mathcal{G}_{P1}$	$\mathcal{G}_{P2}$	$\mathcal{G}_{P3}$	$\mathcal{G}_{P4}$	$\mathcal{G}_{P5}$	$\mathcal{G}_{P6}$	$\mathcal{G}_{P7}$	$\mathcal{G}_{P8}$	$\mathcal{G}_{P9}$	$\mathcal{G}_{P10}$	$\mathcal{G}_{P11}$
$f_1(G_1^*)$	1	1	1	1	1	1	1	1	1	1	1
$f_2(G_2^*)$	0	0	0	1/49	5/105	66/434	131/644	87/589	157/903	104/659	95/717
$f_3(G_3^*)$	0	0	0	0	0	8/434	31/644	15/589	40/903	20/659	10/717
$f_4(G_4^*)$	0	0	0	0	0	0	8/644	0	8/903	2/659	0
$f_5(G_5^*)$	0	0	0	0	0	0	1/644	0	0	0	0
	$\mathcal{G}_{C1}$	$\mathcal{G}_{C2}$	$\mathcal{G}_{C3}$	$\mathcal{G}_{C4}$	$\mathcal{G}_{C5}$	$\mathcal{G}_{C6}$	$\mathcal{G}_{C7}$	$\mathcal{G}_{C8}$	$\mathcal{G}_{C9}$	$\mathcal{G}_{C10}$	$\mathcal{G}_{C11}$
$f_1(G_1^*)$	1	1	1	1	1	1	1	1	1	1	1
$f_2(G_2^*)$	6/140	5/119	11/156	6/128	10/152	14/199	61/361	65/442	71/439	93/592	131/687
$f_3(G_3^*)$	0	0	0	0	0	0	9/361	11/442	8/439	16/592	29/687
$f_4(G_4^*)$	0	0	0	0	0	0	0	0	0	1/592	4/687

number can no longer be associated with any trivial clique or a tiny fraction of the maximum K-core. Both Peter and Carl sequences show that the chromatic number is often close to the clique number (figure 2*a*,*b*). Maximum K^*-core size is generally more than twice the size it would have in the case that it would form a clique—see figure 2 (inset). We therefore conclude that an important part of the whole network structure has enough connectivity to enable the emergence of a non-trivial K-core structure. The whole picture points towards the existence of a broad connectivity pattern responsible for the emergence of increasing chromatic numbers. Nevertheless, even acknowledging the role and suitability of the indicators of validity for the chromatic number used here—maximum K-core, maximum clique number and sequences of *relative energy* of successive colorations of the graph—one cannot completely rule out the existence of a pathological, largely statistically deviated small set of nodes responsible for the behaviour of the chromatic number. We observe that a conclusive response would involve the analysis of the combinatorics among all subsets of the network, which defines a computationally unaffordable problem. We warn the reader that this problem is not restricted to the chromatic number, but it may be present in almost any network measure.

3.3. Real syntax versus null model

Here, we compare the evolution of the chromatic number in real and simulated networks. A data-driven, syntax-free model that generates random child's utterances having the same statistics of word production as Peter and Carl datasets is used as a null model [7]. Underlying the null model outlined below, there is the aim to *reproduce a syntax-free speech flow*, with the same statistical indicators as the real data. That means that we prioritized the simulation of a speaker whose statistics over words usage and sentence length mimic the ones given by the data. Different realizations of the model may lead, for example, to slightly different number of used words—for it is a stochastic phenomena with fluctuations at the sizes we are working in. This is due to the fact that our aim has been, not to randomize the network itself—which would have been the standard approach—but *the process that creates the network*. Since the network is a surrogate of an underlying phenomenon, it is more realistic to create a random version of such underlying phenomenon and, then, build the network, than randomize the network itself. This model definition enables us to assess if the high combinatorics displayed by post-transition networks emerge directly from an increasingly rich vocabulary. We build our model by extracting the following statistical parameters from the 11 recorded conversations in Peter and Carl corpora:

(1) The number of sentences $|S_P(i)|$, $S_C(i)$ in the Peter and Carl datasets.
(2) The probability distribution of *structure lengths* or the probability $P(s)$ that any syntactic structure has s words. We obtain two different distributions, one for each dataset.
(3) We assume that the probability of the ith most frequent word is a scaling law

$$p(i) = \frac{1}{Z} i^{-\beta}, \tag{3.1}$$

with $1 \leq i \leq N_w(T)$, $\beta \approx 1$—i.e. Zipf's Law—and Z is the normalization constant

$$Z = \sum_{i=1}^{N_w(T)} \left(\frac{1}{i}\right)^{\beta}. \tag{3.2}$$

Note that Z depends on lexicon size, $N_w(T)$, which grows slowly at this stage.

We run the above model in the two datasets by generating $|S_{P,C}(i)|$ random sentences, each experiment is repeated 20 times. From the collection of randomly generated syntactic structures we construct a comparable sequence of syntax networks following the same method as in the real datasets (see §3.1). Figure 3 shows that our model generates random syntax networks with size and connectivity comparable to the ones measured in real networks. These statistical indicators display a huge increase during the studied period, this increase being sharper around the age of two, i.e. during the syntactic spurt [7]. As discussed in §2, both the mean connectivity and network size play an important role when determining the values of ω, χ and K^*.

Now, we compute the sequence of averaged chromatic numbers, $\tilde{s}_P(\chi)$, $\tilde{s}_C(\chi)$, for the simulated Peter and Carl syntax networks. Similarly, we generate the sequences of average clique number $\tilde{\Omega}_{P,C}$ and the average maximum K-core $\tilde{\kappa}_{P,C}$. The most salient property we find when comparing real networks obtained from both Peter and Carl's corpora with their randomized counterparts is a huge increase of χ, ω and K^* in the simulated networks. That is, the ensemble of random strings displays

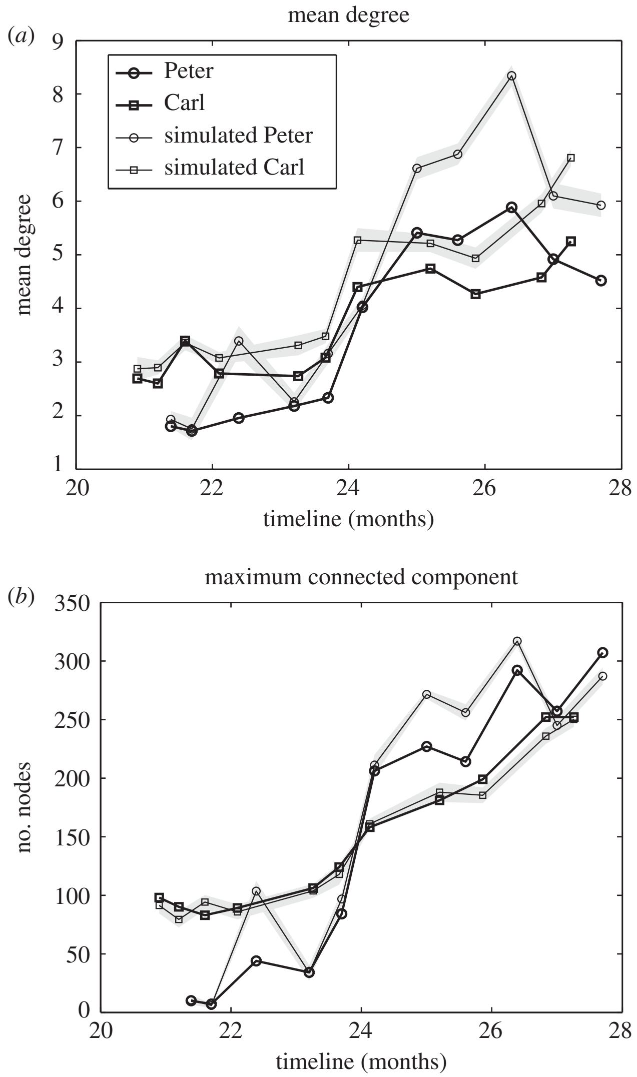

Figure 3. Evolution of (*a*) the mean degree and (*b*) size of the largest connected component in the real (strong solid lines) and simulated (weak solid lines) syntax networks. Shaded grey areas correspond to standard deviation in the case of the simulated instances.

higher complexity parameters than the real corpora. For example, at the end of the studied period, the three complexity estimators are close to 10 in Peter simulations and close to 9 in Carl simulations (figure 2*c*,*d*).

A very interesting feature is found at the first stages of the simulated Peter sequence: the random networks are no longer bipartite—see §3.2. In particular, the third random corpus has an average chromatic number of 4, which is significantly higher than the observed chromatic number. In this case, the two-stage grammar imposes severe constraints on what is actually plausible in any pre-transition syntactic structure. This trend is also observed at later stages of language acquisition. In general, simulated networks have higher chromatic numbers than empirical networks, although both two types of networks have similar connectivities—by definition. In some cases, the average chromatic number of the graphs belonging to the random ensemble is twice the real one (figure 2). To better understand the nature of these deviations, we have compared the behaviour of chromatic numbers against mean connectivity and the size of the largest connected component. Figure 4 shows a well-defined, non-trivial deviation between real networks and random networks. In particular, when comparing the relation between the chromatic number and the average degree of Peter's corpus (figure 4*a*) and Carl's corpus (figure 4*b*) with the simulated ones, we observe a clear trend of the real networks towards smaller chromatic numbers. The comparison of the size of the giant connected component, *GCC*, clearly depending on the size in the case of scale-free networks [29], shows the same trend both in Peter's (figure 4*c*) and Carl's (figure 4*d*) corpora. In general, the expected

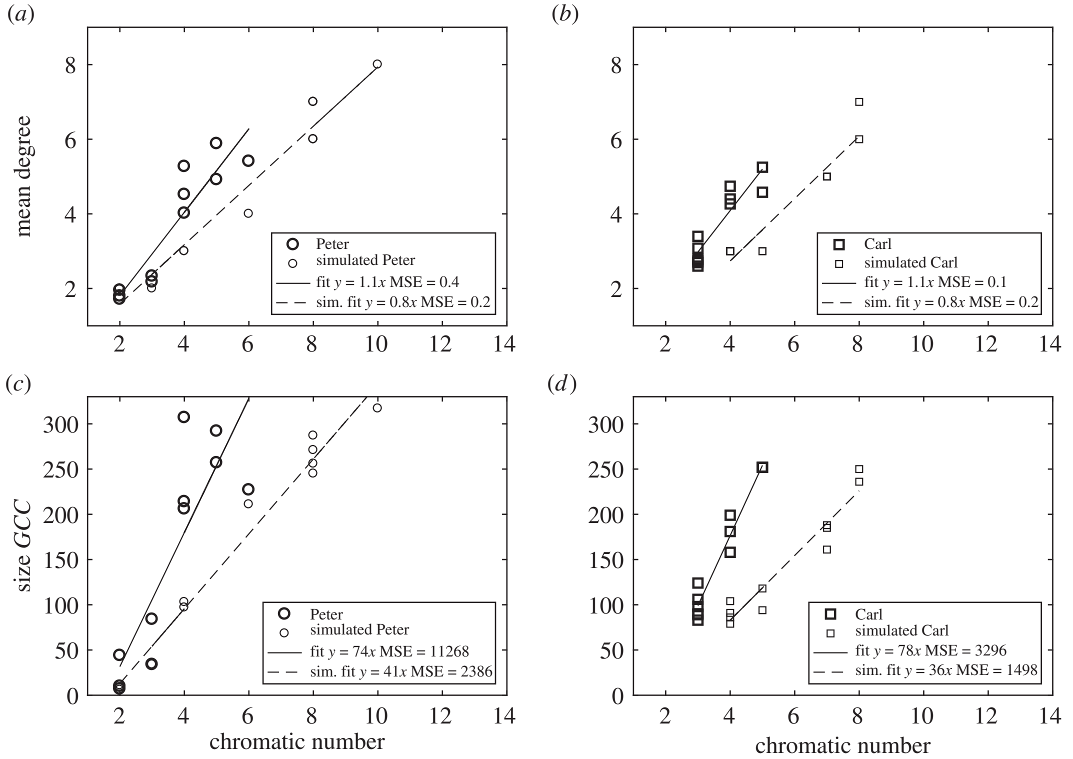

Figure 4. Relationship between the chromatic number, mean degree and giant connected component (*GCC*), comparing real data from the null model. (*a*) Scatter plot between the chromatic number and the average connectivity of the networks for Peter's corpus and its associated null model. (*b*) Scatter plot between the chromatic number and the average connectivity of the networks for Carl's corpus and its associated null model. (*c*) Scatter plot between the chromatic number and the size of the *GCC* for Peter's corpus and its associated null model. (*d*) Scatter plot between the chromatic number and the size of the *GCC* for Carl's corpus and its associated null model. Clearly, all indicators show that the chromatic number observed in real graphs is below that expected by a non-syntactic speaker.

chromatic number is larger than the one observed in the real networks. These plots suggest that the chromatic number is capturing essential combinatorial properties of the underlying system, which cannot be reproduced with a simple, syntax-free random generation model. We support this argument by providing a linear regression fit to the data. In the case of the relation of the *GCC* and the chromatic number, the mean squared error (MSE) is substantially higher in the case of the real instances, Peter and Carl (figure 4).

4. Discussion

Syntax is a characteristic, complex and defining feature of language organization. It pervades its capacity for unbounded generative power of the linguistic system [5], allows sentences to be organized in highly structured ways and is acquired in almost full power by children after being exposed to a limited repertoire of examples. Syntax is also one aspect of the whole: semantic and phonological aspects need to be taken into account, and they are all embedded in (and run by) a cognitive, brain-embodied framework [35]. Because of the dominant role played by how words actually interact with each other, computational and theoretical approaches dealing with word inventories or other statistical trends ignoring interactions are likely to be limited. As an example of the high degree of intricacy involved in linguistic acquisition, we mention two recent works: First, recent studies on acquisition in French toddlers provided strong evidences for non-trivial interactions between phonological, semantic and syntactic *modules*, showing the presence of inhibition/activation patterns in the acquisition dynamics involving cross-dependencies among those modules [36]. The second example comes from the framework of *multiplex networks*—i.e. networks involving different layers of interaction (see [37,38] and references therein). Specifically, it has been shown that, taking into account different layers of interactions, a critical phase can be identified from a simple, restricted grammar towards a flexible, more complex grammar [18,19], consistent with the results provided in our paper.

Following previous work that takes advantage of complex networks approaches to language organization [13] we have made a step further in studying the structure of syntax graphs using graph colouring. The motivation of this approximation is twofold. On the one hand, graph colourability allows to properly detect correlations that are not captured by topological approaches. On the other hand, it seems a natural way to substantiate previous claims connecting syntax with compatibility relations common with other types of systems, such as chemical structures. In this context, standard network measurements like average degree, clustering or degree distribution are much more limited. Since graph colouring naturally defines compatibility through the presence or absence of a common label to every pair of nodes, it seems the right framework to study the process of network growth in child language. The behaviour of the chromatic number accurately marks the syntactic spurt in language acquisition, i.e. it is a footprint of the generative power of the underlying grammar.

There are limitations associated with the network definition. Syntactic relations are structure-dependent, not sequence dependent. Because the network is an aggregation of text sequences, it cannot fully grasp the hierarchical nature associated with syntactic constructs. Still, the chromatic number is a global measurement that can detect grammar constraints by analysing the pattern of network interaction at different scales. That is, the network representation is an indicator of global linguistic proficiency and includes some combinatorial signal which can be properly detected with the chromatic number. Besides the suitability of the measure, it is in force to highlight that more longitudinal studies are needed. In this case, we studied two single individuals whose data is of excellent quality. Moreover, we assembled the syntactic networks by discussing the linguistic validity of each syntactic relation in detail. We therefore have chosen this high level of accuracy in our analysis, in spite of performing a massive one with less delicate assembling methodology. Further studies should perform much more longitudinal explorations, involving eventually other languages or bilingual/multilingual children, with the same degree of detail in the analysis, when possible.

There are other, broader implications of our work. The chromatic number can be viewed as a reciprocal measure of standard community detection. Here, the chromatic number defines a partition of the network in classes of unlinked nodes. This definition is particularly relevant in networks where some kind of compatibility relation is at work in the wiring process. In this case, the standard community structure can be misleading, because elements of the same class cannot be connected. The case for syntactic graphs is paradigmatic but the partition induced by the chromatic number could shed light into the behaviour of many other systems. Additionally, we have proposed to assess the statistical significance of these partitions with the sequence of minimal violations—see equation (2.10). Future work should explore how the chromatic number—and related measures—can be exploited to detect *forbidden* links in the network. Deviations of the chromatic number, as the ones observed in this paper, suggest the presence of combinatorial constraints that must be taken into account, for example, when defining proper null-models.

Data accessibility. Data have been extracted from the CHILDES database. Transcripts from Peter's corpus are found in: https://childes.talkbank.org/browser/index.php?url=Eng-NA/Bloom70/Peter/. Transcripts for Carl's corpus are found in: https://childes.talkbank.org/browser/index.php?url=Eng-UK/Manchester/carl/.

Authors' contributions. B.C.-M., M.S.F., S.V. and R.S. equally contributed to the idea, development of the research and writing of the manuscript. B.C.-M. performed the linguistic analysis and M.S.F. and S.V. performed the numerical analysis. R.S. conducted the research.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by the James McDonnell Foundation (B.C-M., S.V. and R.S.) and the Spanish Plan NacionalINSOCO DPI2016-80116-P (M.S.F.).

Acknowledgements. We thank Complex Systems Laboratory members for fruitful conversations.

References

1. Maynard-Smith J, Szathmàry E. 1997 *The major transitions in evolution*. New York, NY: University of New York Press.
2. Bickerton D. 1990 *Language and species*. Chicago: University of Chicago Press.
3. Hauser MD, Chomsky N, Fitch TW. 2002 The faculty of language: what is it, who has it, and how did it evolve? *Science* **298**, 1569–1579. (doi:10.1126/science.298.5598.1569)
4. Christiansen MH, Kirby S. 2003 Language evolution: consensus and controversies. *Trends Cogn. Sci.* **7**, 300–307. (doi:10.1016/S1364-6613(03)00136-0)
5. Chomsky N. 1988 *Language and problems of knowledge*. Cambridge, MA: MIT Press.
6. Radford A. 1990 *Syntactic theory and the acquisition of english syntax: the nature of early child grammars of English*. Oxford, UK: Blackwell.
7. Corominas-Murtra B, Valverde S, Solé R. 2009 The ontogeny of scale-free syntax networks: phase transitions in early language acquisition. *Adv. Complex Syst. (ACS)* **12**, 371–392. (doi:10.1142/S0219525909002192)
8. Barceló-Coblijn L, Corominas-Murtra B, Gomila A. 2012 Syntactic trees and small-world networks: syntactic development as a dynamical

process. *Adapt. Behav.* **20**, 427–442. (doi:10.1177/1059712312455439)

9. Blythe RA, McKane AJ. 2007 Stochastic models of evolution in genetics, ecology and linguistics. *J. Stat. Mech: Theory Exp.* **2007**, P07018. (doi:10.1088/1742-5468/2007/07/p07018)
10. Solé RV, Corominas-Murtra B, Fortuny J. 2010 Diversity, competition, extinction: the ecophysics of language change. *J. R. Soc. Interface* **7**, 1647–1664. (doi:10.1098/rsif.2010.0110)
11. Ferrer-i-Cancho R, Köhler R, Solé RV. 2004 Patterns in syntactic dependency networks. *Phys. Rev. E* **69**, 051915. (doi:10.1103/PhysRevE.69.051915)
12. Steele JL. 2009 A hubterranean view of syntax: an analysis of linguistic form through network theory. PhD thesis, Doctoral dissertation, University of Queensland, Australia, p. 305.
13. Solé RV, Corominas-Murtra B, Valverde S, Steels L. 2010 Genome size, self-organization and DNA's dark matter. *Complexity* **15**, 20–23. (doi:10.1002/cplx.20326)
14. Baronchelli A, Ferrer-i Cancho R, Pastor-Satorras R, Chater N, Christiansen MH. 2013 Networks in cognitive science. *Trends Cogn. Sci.* **17**, 348–360. (doi:10.1016/j.tics.2013.04.010)
15. Karuza EA, Thompson-Schill SL, Bassett DS. 2016 Local patterns to global architectures: influences of network topology on human learning. *Trends Cogn. Sci.* **20**, 629–640. (doi:10.1016/j.tics.2016.06.003)
16. Ke JY, Yao. Y. 2008 Analysing language development from a network approach. *J. Quant. Linguist.* **15**, 70–99. (doi:10.1080/09296170701794286)
17. Beckage N, Smith L, Hills T. 2011 Small worlds and semantic network growth in typical and late talkers. *PLoS ONE* **6**, e19348. (doi:10.1371/journal.pone.0019348)
18. Stella M, Beckage NM, Brede M. 2017 Multiplex lexical networks reveal patterns in early word acquisition in children. *Sci. Rep.* **7**, 46730. (doi:10.1038/srep46730)
19. Stella M, Beckage NM, Brede M, DeDomenico M. 2018 Multiplex model of mental lexicon reveals explosive learning in humans. *Sci. Rep.* **8**, 2259. (doi:10.1038/s41598-018-20730-5)
20. Tauber KRSJ. 1971 Linguistics as a basis for analyzing chemical structure diagrams. *J. Chem. Doc* **11**, 139–141. (doi:10.1021/c160042a005)
21. Brooks RI, Tutte WT. 1941 On colouring the nodes of a network. *Proc. Camb. Phil. Soc.* **39**, 194–197. (doi:10.1017/s030500410002168x)
22. Bollobás B. 1998 *Modern graph theory*, corrected edn. Berlin, Germany: Springer.
23. Bollobás B. 2001 *Random graphs*. Cambridge, UK: Cambridge University Press.
24. Wu FY. 1982 The Potts model. *Rev. Mod. Phys.* **54**, 235–268. (doi:10.1103/RevModPhys.54.235)
25. Bollobás B. 1988 The chromatic number of random graphs. *Comb* **8**, 49–55. (doi:10.1007/bf02122551)
26. Achlioptas D, Molloy M. 1999 Almost all graphs with 2.522*n* edges are not 3-colourable. *Electron. J. Comb.* **6**, R29. See http://www.combinatorics.org/ojs/index.php/eljc/article/view/v6i1r29.
27. Zdeborová L, Krzakala F. 2007 Phase transitions in the coloring of random graphs. *Phys. Rev. E* **76**, 031131. (doi:10.1103/PhysRevE.76.031131)
28. Melçuck I. 1989 *Dependency grammar: theory and practice*. New York, NY: Oxford University Press.
29. Bianconi G, Marsili M. 2006 Emergence of large cliques in random scale-free networks. *EPL (Europhysics Letters)* **74**, 740–746. (doi:10.1209/epl/i2005-10574-3)
30. Corominas-Murtra B. 2007 Network statistics on early English syntax: structural criteria. (http://arxiv.org/abs/0704.3708)
31. Corominas-Murtra B. 2011 A unified approach to the emergence of complex communication. PhD dissertation, Universitat Pompeu Fabra, Barcelona, Spain.
32. Hristea F, Popescu M. 2003 *Dependency grammar annotator, building awareness in language technology*. Bucarest, Romania: Editura Universitatii din Bucaresti.
33. Bloom L, Hood L, Lightbown P. 1974 Imitation in language development: if, when, and why. *Cognit. Psychol.* **6**, 380–420. (doi:10.1016/0010-0285(74)90018-8)
34. Bloom L, Lightbown P, Hood L. 1975 *Structure and variation in child language.* Monographs of the Society for Research in Child Development, Serial 160.
35. Jackendoff R. 2002 *Foundations of language: brain, meaning, grammar, evolution*. Oxford, UK: Oxford University Press.
36. Dautriche I, Swingley D, Christophe A. 2015 Learning novel phonological neighbors: syntactic category matters. *Cognition* **143**, 77–86. (doi:10.1016/j.cognition.2015.06.003)
37. Kivela M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter M. 2014 Multilayer networks. *J. Complex Netw.* **2**, 203–271. (doi:10.1093/comnet/cnu016)
38. Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S. 2010 Catastrophic cascade of failures in interdependent networks. *Nature* **464**, 1025–1028. (doi:10.1038/nature08932)