Biosemiotics (2021) 14:345–375
https://doi.org/10.1007/s12304-021-09452-w

ORIGINAL RESEARCH

The Advent and Fall of a Vocabulary Learning Bias from Communicative Efficiency

David Carrera-Casado[1] · **Ramon Ferrer-i-Cancho**[1]

Received: 27 May 2021 / Accepted: 10 September 2021 / Published online: 26 November 2021

Abstract

Biosemiosis is a process of choice-making between simultaneously alternative options. It is well-known that, when sufficiently young children encounter a new word, they tend to interpret it as pointing to a meaning that does not have a word yet in their lexicon rather than to a meaning that already has a word attached. In previous research, the strategy was shown to be optimal from an information theoretic standpoint. In that framework, interpretation is hypothesized to be driven by the minimization of a cost function: the option of least communication cost is chosen. However, the information theoretic model employed in that research neither explains the weakening of that vocabulary learning bias in older children or polylinguals nor reproduces Zipf's meaning-frequency law, namely the non-linear relationship between the number of meanings of a word and its frequency. Here we consider a generalization of the model that is channeled to reproduce that law. The analysis of the new model reveals regions of the phase space where the bias disappears consistently with the weakening or loss of the bias in older children or polylinguals. The model is abstract enough to support future research on other levels of life that are relevant to biosemiotics. In the deep learning era, the model is a transparent low-dimensional tool for future experimental research and illustrates the predictive power of a theoretical framework originally designed to shed light on the origins of Zipf's rank-frequency law.

Keywords Biosemiosis · Vocabulary learning · Mutual exclusivity · Zipfian laws · Information theory · Quantitative linguistics

✉ David Carrera-Casado
david.carrera@estudiantat.upc.edu

Ramon Ferrer-i-Cancho
rferrericancho@cs.upc.edu

[1] Complexity and Quantitative Linguistics Lab, LARCA Research Group, Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona, Catalonia, Spain

Introduction

Biosemiotics can be defined as a science of signs in living systems (Kull, 1999, p. 386). Here we join the effort of developing such a science. Focusing on the problem of "learning" new signs, we hope to contribute (i) to place choice at the core of semiotic theory of learning (Kull, 2018) and (ii) to make biosemiotics compatible with the information theoretic perspective that is regarded as currently dominant in physics, chemistry, and molecular biology (Deacon, 2015).

Languages use words to convey information. From a semantic perspective, words stand for meanings (Fromkin et al., 2014). Correlates of word meaning have been investigated in other species (e.g. Hobaiter & Byrne, 2014; Genty & Zuberbühler, 2014; Moore, 2014). From a neurobiological perspective, words can be seen as the counterparts of cell assemblies with distinct cortical topographies Pulvermüller (2001, 2013).

From a formal standpoint, the essence of that research is some binding between a sign or a form, e.g., a word or an ape gesture, and a counterpart, e.g. a 'meaning' or an assembly of cortical cells. Mathematically, that binding can be formalized as a bipartite graph where vertices are forms and their counterparts (Fig. 1). Such abstract setting allows for a powerful exploration of natural systems across levels of life, from the mapping of animal vocal or gestural behaviors (Fig. 2a) into their "meanings" down to the mapping from codons into amino acids (Fig. 2b) while allowing for a comparison against "artificial" coding systems such as the Morse code (Fig. 2c) or those emerging in artificial naming games (Hurford, 1989; Steels, 1996). In that setting, almost connectedness has been hypothesized to be the mathematical condition required for the emergence of a rudimentary form of syntax and symbolic reference (Ferrer-Cancho et al., 2005, 2006). By symbolic reference, we mean here Deacon's revision of Pierce's view (Deacon, 1997). The almost connectedness condition is met when it is possible to reach practically any other vertex of the network by starting a walk from any possible vertex (as in Fig. 1a and b but not in Fig. 1c and d).

Since the pioneering research of G. K. Zipf (1949), statistical laws of language have been interpreted as manifestations of the minimization of cognitive costs (Zipf, 1949; Ellis and Hitchcock, 1986; Gustison et al., 2016; Ferrer-i-Cancho & Díaz-Guilera, 2007, 2019). Zipf argued that the law of abbreviation, the tendency of more frequent words to be shorter, resulted from a minimization of a cost function involving, for every word, its frequency, its "mass" and its "distance", which in turn implies the minimization of the size of words (Zipf, 1949, p.59). Recently, it as been shown mathematically that the minimization of the average of the length of words (the mean code length in the language of information theory) predicts a correlation between frequency and duration that cannot be positive, extending and generalizing previous results from information theory (Ferrer-i-Cancho et al., 2019). The framework addresses the general problem of assigning codes as short as possible to counterparts represented by distinct numbers while warranting certain constraints, e.g., that every number will receive a distinct code (e.g. non-singular coding in the language of information theory). If the counterparts are word types from a vocabulary, it predicts the law of abbreviation as it occurs in the vast majority of languages (Bentz & Ferrer-i-Cancho, 2016). If these counterparts are meanings, it predicts that more frequent

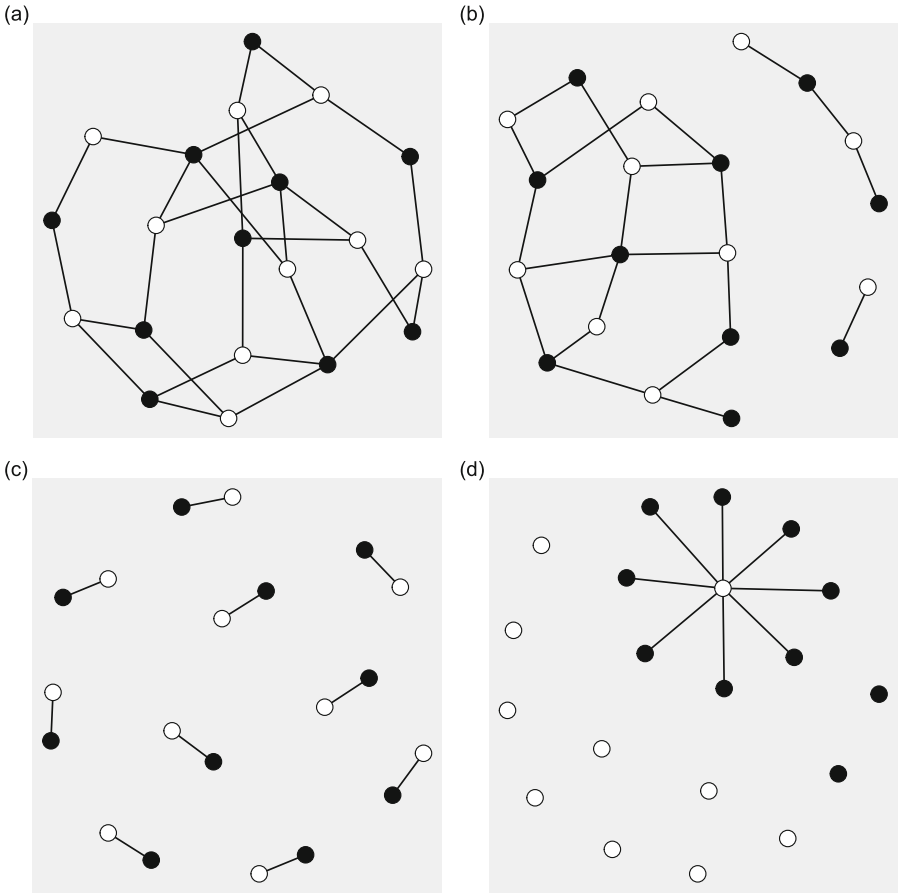

Fig. 1 A bipartite graph linking forms (white circles) with their counterparts (black circles). **a** a connected graph **b** an almost connected graph **c** a one-to-one mapping between forms and counterparts **d** a mapping where only one form is linked with counterparts

meanings should tend to be assigned smaller codes (e.g., shorter words) as found in real experiments (Kanwal et al., 2017; Brochhagen, 2021). Table 1 summarizes these and other predictions of compression.

A family of probabilistic models

The bipartite graph of form-counterpart associations is the *skeleton* (Figs. 1 and 2) on which a family of models of communication has been built (Ferrer-i-Cancho & Díaz-Guilera 2007, 2018). The target of the first of these models (Ferrer-i-Cancho & Sole, 2003) was Zipf's rank-frequency law, that defines the relationship between the frequency of a word f and its rank i, approximately as

$$f \approx i^{-\alpha}.$$

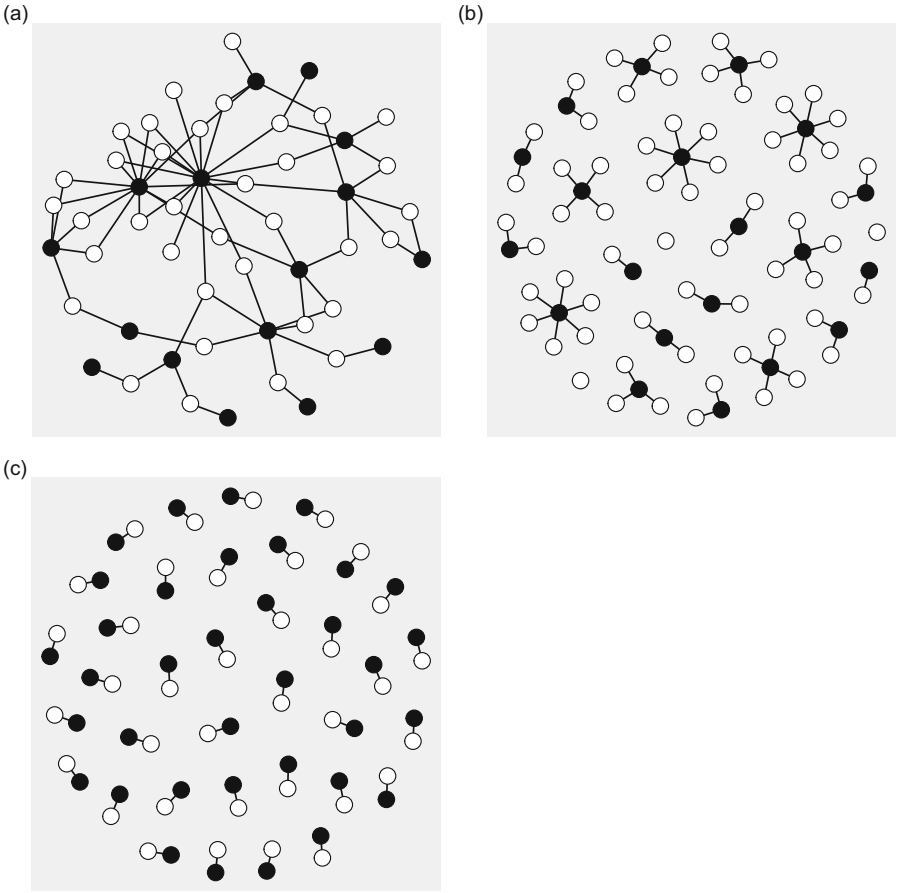

Fig. 2 Real bipartite graphs linking forms (white circles) with their counterparts (black circles). **a** Chimpanzee gestures and their meaning (Hobaiter & Byrne, 2014, Table S3). This table was chosen for its broad coverage of gesture types (see other tables satisfying other constraints, e.g. only gesture-meaning associations employed by a sufficiently large number of individuals). **b** Codon translation into amino acids, where forms are 64 codons and counterparts are 20 amino acids **c** The international Morse code, where forms are strings of dots and dashed and the counterparts are letters of the English alphabet (A, B, ..., Z) and digits (0,1,...,9)

These early models were aimed at shedding light on mainly three questions:

1. The origins of this law (Ferrer-i-Cancho & Sole, 2003, 2005b).
2. The range of variation of α in human language (Ferrer & Cancho, 2005a, 2006).
3. The relationship between α and the syntactic and referential complexity of a communication system (Ferrer-i-Cancho et al., 2005, 2006).

The main assumption of these models is that word frequency is an epiphenomenon of the structure of the skeleton or the probability of the meanings. Following the metaphor of the skeleton, the models are *bodies* whose *flesh* are probabilities that are calculated from the skeleton. The first models defined $p(s_i|r_j)$, the probability that

Table 1 The application of the scientific method in quantitative linguistics (italics) with various concrete examples (roman)

linguistic laws	⟶ *principles*	⟶ *predictions*
		Köhler (1987) and Altmann (1993)
Zipf's law of abbreviation	⟶ compression	⟶ Menzerath's law (Gustison et al., 2016; Ferrer-i-Cancho et al., 2019)
		⟶ Zipf's rank-frequency law Ferrer-i-Cancho (2016a)
		⟶ "shorter words" for more frequent "meanings" (Ferrer-i-Cancho et al., 2019; Kanwal et al., 2017; Brochhagen, 2021)
Zipf's rank-frequency law	⟶ mutual information maximization + surprisal minimization	⟶ **a vocabulary learning bias** Ferrer-i-Cancho (2017a)
		⟶ the principle of contrast Ferrer-i-Cancho (2017a)
		⟶ range or variation of α Ferrer-Cancho (2005a, 2006)

α is the exponent of Zipf's rank-frequency law (Zipf, 1949). The prediction that is the target of the current article is shown in boldface

a speaker produces s_i given a counterpart r_j, as the same for all words connected to r_j. In the language of mathematics,

$$p(s_i|r_j) = \frac{a_{ij}}{\omega_j}, \tag{1}$$

where a_{ij} is a boolean (0 or 1) that indicates if s_i and r_j are connected and ω_j is the degree of r_j, namely the number of connections of r_j with forms, i.e.

$$\omega_j = \sum_i a_{ij}.$$

These models are often portrayed as *models of the assignment of meanings to forms* (Futrell, 2020; Piantadosi, 2014) but this description falls short because

- They are indeed models of production as they define the probability of producing a form given some counterparts (as in Eq. 1) or simply the marginal probability of a form. The claim that *theories of language production or discourse do not explain the law* (Piantadosi, 2014) has no basis and raises the questions of which theories of language production are deemed acceptable.
- They are also models of understanding, as they define symmetric conditional probabilities such as $p(r_j|s_i)$, the probability that a listener interprets r_j when receiving s_i.
- The models are flexible. In addition to "meaning", other counterparts were deemed possible from their birth. See for instance the use of the term "stimuli" (Ferrer-i-Cancho & Díaz-Guilera, 2007, e.g.), as a replacement for meaning that was borrowed from neurolinguistics (Pulvermüller, 2001).
- The models fit in the distributional semantics framework (Lund & Burgess, 1996) for two reasons: their flexibility, as counterparts can be dimensions in some hidden space, and also because of representing a form as a vector of their joint or conditional probabilities with "counterparts" that is inferred from the network structure, as we have already explained (Ferrer-i-Cancho & Vitevitch, 2018).

Contrary to the conclusions of Piantadosi (2014), there are derivations of Zipf's law that do account for psychological processes of word production, especially the intentionality of choosing words in order to convey a desired meaning.

The family of models assume that the skeleton that determines all the probabilities, the bipartite graph, is shaped by a combination of minimization of the entropy (or surprisal) of words (H) and the maximization of the mutual information between words and meanings (I), two principles that are cognitively motivated and that capture speaker and listener's requirements (Ferrer-i-Cancho, 2018). When only the entropy of words is minimized, configurations where only one form is linked as in Fig. 1d are predicted. When only the mutual information between forms and counterparts is maximized, one-to-one mappings between forms and counterparts are predicted (when the number of forms and counterparts is the same) as in Figs. 1c or 2d. Real language is argued to be in-between these two extreme configurations (Ferrer-i-Cancho & Díaz-Guilera, 2007). Such a trade-off between simplicity (Zipf's unification) and effective communication (Zipf's diversification) is also found in information theoretic models of communication based on the information bottleneck approach (see Zaslavsky et al. (2021) and references there in).

In quantitative linguistics, scientific theory is not possible without taking into consideration language laws (Köhler, 1987; Debowski, 2020). Laws are seen as manifestations of principles (also referred as *"requirements"* by Köhler 1987), which are key components of explanations of linguistic phenomena. As part of the scientific method cycle, novel predictions are key aim (Altmann, 1993) and key to validation and refinement of theory (Bunge, 2001). Table 1 synthesizes this general view as chains of the form: *laws*, *principles* that are inferred from them, and *predictions* that are made from those principles, giving concrete examples from previous research.

Although one of the initial goals of the family of models was to shed light on the origins of Zipf's law for word frequencies, a member of the family of models

turned out to generate a novel prediction on vocabulary learning in children and the tendency of words to contrast in meaning (Ferrer-i-Cancho, 2017a): when encountering a new word, children tend to infer that it refers to a concept that does not have a word attached to it (Markman & Wachtel, 1988; Merriman & Bowman, 1989; Clark, 1993). The finding is cross-linguistically robust: it has been found in children speaking English (Markman & Wachtel, 1988), Canadian French (Nicoladis & Laurent, 2020), Japanese (Haryu, 1991), Mandarin Chinese (Byers-Heinlein & Werker, 2013; Hung et al., 2015), Korean (Eun-Nam, 2017). These languages correspond to four distinct linguistic families (Indo-European, Japonic, Sino-Tibetan, Koreanic). Furthermore, the finding has also been replicated in adults (Hendrickson & Perfors, 2019; Yurovsky & Yu, 2008) and other species (Kaminski et al., 2004). This phenomenon is a example of biosemiosis, namely a process of choice-making between simultaneously alternative options (Kull, 2018, p. 454).

As an explanation for vocabulary learning, the information theoretic model suffers from some limitations that motivate the present article. The first one is that the vocabulary learning bias weakens in older children (Kalashnikova et al., 2016; Yildiz, 2020) or in polylinguals (Houston-Price et al., 2010; Kalashnikova et al., 2015), while the current version of the model predicts the vocabulary learning bias only provided that mutual information maximization is not neglected (Ferrer-i-Cancho, 2017a).

The second limitation is inherited from the family of models, where the definition of the probabilities over the bipartite graph skeleton leads to a linear relationship between the frequency of a form and its number of counterparts (Ferrer-i-Cancho & Vitevitch, 2018). However, this is inconsistent with Zipf's prediction, namely that the number of meanings μ a word of frequency f should follow (Zipf, 1945)

$$\mu \approx f^{\delta}, \tag{2}$$

with $\delta = 0.5$. Equation 2 is known as Zipf's meaning-frequency law (Zipf, 1949). To overcome such a limitation, Ferrer-i-Cancho and Vitevitch (2018) proposed different ways of modifying the definition of the probabilities from the skeleton. Here we borrow a proposal of defining the joint probability of a form and its counterpart as

$$p(s_i, r_j) \propto a_{ij}(\mu_i \omega_j)^{\phi}, \tag{3}$$

where ϕ is a parameter of the model and μ_i and ω_j are, respectively, the degree (number of connections) of the form s_i and the counterpart r_j. Previous research on vocabulary learning in children with these models (Ferrer-i-Cancho, 2017a) assumed $\phi = 0$, which leads to $\delta = 1$ (Ferrer-i-Cancho, 2016b). When $\phi = 1$, the system is channeled to reproduce Zipf's meaning-frequency law, i.e. Equation 2 with $\delta = 0.5$ (Ferrer-i-Cancho & Vitevitch, 2018).

Overview of the present article

It has been argued that there cannot be meaning without interpretation (Eco, 1986). As Kull (2020) puts it, "*Interpretation (which is the same as primitive decision-making) assumes that there exists a choice between two or more options. The options can be described as different codes applicable simultaneously in the same situation.*" The main aim to of this article is to shed light on the choice between strategy a, i.e.

attaching the new form to a counterpart that is unlinked, and strategy b, i.e. attaching the new form to a counterpart that is already linked (Fig. 3).

The remainder of the article is organized as follows. Section "The Mathematical Model" considers a model of a communication system that has three components:

1. A *skeleton* that is defined by a binary matrix A that indicates the form-counterpart connections.
2. A *flesh* that is defined over the skeleton with Eq. 3,
3. A *cost function*, that defines the cost of communication as

$$\Omega = -\lambda I + (1 - \lambda)H, \tag{4}$$

where λ is a parameter that regulates the weight of mutual information (I) maximization and word entropy (H) minimization such that $0 \leq \lambda \leq 1$. I and H are inferred from matrix A and Eq. 3 (further details are given in "The Mathematical Model").

This section introduces Δ, i.e. the difference in the cost of communication between strategy a and strategy b according to Ω (Fig. 3). $\Delta < 0$ indicates that the cost of communication of strategy a is lower than that of b. Our main hypothesis is that interpretation is driven by the Ω cost function and that a receiver will choose the option that minimizes the resulting Ω. By doing this, we are challenging the longstanding and limiting belief that information theory is dissociated from semiotics and not concerned about meaning (Deacon, 2015, e.g.). This article is a just one counterexample (see also Zaslavsky et al. (2018)). Information theory, as any abstract powerful mathematical tool, can serve applications that do not assume meaning (or meaning-making processes) as in the original setting of telecommunication where it was developed by Shannon, as well as others that do, although they were not his primary concern for historical and sociological reasons.

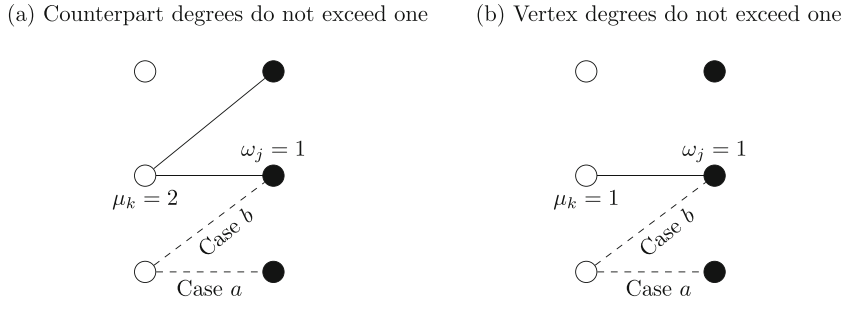

Fig. 3 Strategies for linking a new word to a meaning. Strategy a consists of linking a word to a free meaning, namely an unlinked meaning. Strategy b consists of linking a word to a meaning that is already linked. We assume that the meaning that is already linked is connected to a single word of degree μ_k. Two simplifying assumptions are considered. **a** Counterpart degrees do not exceed one, implying $\mu_k \geq 1$. **b** Vertex degrees do not exceed one, implying $\mu_k = 1$

In general, the formula of Δ is complex and the analysis of the conditions where a is advantageous (namely $\Delta < 0$) requires making some simplifying assumptions. If $\phi = 0$, then one obtains that Ferrer-i-Cancho (2017a)

$$\Delta = -\lambda \frac{(\omega_j + 1)\log(\omega_j + 1) - \omega_j \log(\omega_j)}{M + 1}, \tag{5}$$

where M is the number of edges in the skeleton and ω_j is the degree of the already linked counterpart that is selected in strategy b (Fig. 3). Equation 5 indicates that strategy a will be advantageous provided that mutual information maximization matters (i.e. $\lambda > 0$) and its advantage will increase as mutual information maximization becomes more important (i.e. for larger λ), the linked counterpart has more connections (i.e. larger ω_j) or when the skeleton has less connections (i.e. smaller M). To be able to analyze the case $\phi > 0$, we will examine two classes of skeleta that are presented next.

Counterpart degrees do not exceed one In this class, the degrees of counterparts are restricted to not exceed one, namely a counterpart can only be disconnected or connected to just one form. If meanings are taken as counterparts, this class matches the view that *"no two words ever have exactly the same meaning"* (Fromkin et al., 2014, p. 256), based on the notion of absolute synonymy (Dangli & Abazaj, 2009).

This class also mirrors the linguistic principle that any two words should contrast in meaning (Clark, 1987). Alternatively, if synonyms are deemed real to some extent, this class may capture early stages of language development in children or early stages in the evolution of languages where synonyms have not been learned or developed. From a theoretical standpoint, this class is required by the maximization of the mutual information between forms and counterparts when the number of forms does not exceed that of counterparts (Ferrer-i-Cancho & Vitevitch, 2018).

We use μ_k to refer to degree of the word that will be connected to meaning selected in strategy b (Fig. 3). We will show that, in this class, Δ is determined by λ, ϕ, μ_k and the degree distribution of forms, namely the vector of form degrees $\vec{\mu} = (\mu_1, ..., \mu_i, ...\mu_n)$.

Vertex degrees do not exceed one In this class, the degrees of any vertex are restricted to not exceed one, namely a form (or a meaning) can only be disconnected or connected to just one counterpart (just one form). This class is narrower than the previous one because it imposes that degrees do not exceed one both for forms and counterparts. Words lack homonymy (or polysemy). We believe that this class would correspond to even earlier stages of language development in children (where children have learned at most one meaning of a word) or earlier stages in the evolution of languages (where the communication system has not developed any homonymy). From a theoretical stand point, that class is a requirement of maximizing mutual information between forms and counterparts when $n = m$ (Ferrer-i-Cancho & Vitevitch, 2018). We will show that Δ is determined just by λ, ϕ and M, the number of links of the bipartite skeleton.

Notice that meanings with synonyms have been found in chimpanzee gestures (Hobaiter & Byrne, 2014), which suggests that the two classes above do not capture

the current state of the development of form-counterpart mappings in adults of other species. Section "The Mathematical Model" presents the formulae of Δ for each classes. Section "Results" uses this formulae to explore the conditions that determine when strategy a is more advantageous, namely $\Delta < 0$, for each of the two classes of skeleta above, that correspond to different stages of the development of language in children. While the condition $\phi = 0$ implies that strategy a is always advantageous when $\lambda > 0$, we find regions of the space of parameters where this is not the case when $\phi > 0$ and $\lambda > 0$. In the more restrictive class, where vertex degrees do not exceed one, we find a region where a is not advantageous when λ is sufficiently small and M is sufficiently large. The size of that region increases as ϕ increases. From a complementary perspective, we find a region where a is not advantageous ($\Delta \geq 0$) when λ is sufficiency small and ϕ is sufficiently large; the size of the region increases as M increases. As M is expected to be larger in older children or in polylinguals (if the forms of each language are mixed in the same skeleton), the model predicts the weakening of the bias in older children and polylinguals (Liittschwager & Markman, 1994; Yildiz, 2020; Houston-Price et al., 2010; Kalashnikova et al., 2015, 2016, 2019). To ease the exploration of the phase space for the class where the degrees of counterparts do not exceed one, we will assume that word frequencies follow Zipf's rank-frequency law. Again, regions where a is not advantageous ($\Delta \geq 0$) also appear but the conditions for the emergence of this regions are more complex. Our preliminary analyses suggest that the bias should weaken in older children even for this class. Section "Discussion" reviews the findings, suggests future research directions and reviews the research program in light of the scientific method.

The Mathematical Model

Below we give more details about the model that we use to investigate the learning of new words and outlines the arguments that take from Eq. 3 to concrete formulae of Δ. Section "Δ in two Classes of Skeleta" just presents the concrete formulae Δ for each of the two classes of skeleta. Full details are given in the Supplementary Information, Section "S1 The mathematical model in detail". The model has four components that we review next.

Skeleton ($A = a_{ij}$) A bipartite graph that defines the associations between n forms and m counterparts that are defined by an adjacency matrix $A = \{a_{ij}\}$.

Flesh ($p(s_i, r_j)$) The flesh consist of a definition of $p(s_i, r_j)$, the joint probability of a form (or word) and a counterpart (or meaning) and a series of probability definitions stemming from it. Probabilities depart from previous work (Ferrer-i-Cancho & Sole, 2003, 2005b) by the addition of the parameter ϕ. Equation 3 defines $p(s_i, r_j)$ as proportional to the product of the degrees of the form and the counterpart to the power of ϕ, which is a parameter of the model. By normalization, namely

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p(s_i, r_j) = 1,$$

Equation 3 leads to

$$p(s_i, r_j) = \frac{1}{M_\phi} a_{ij} (\mu_i \omega_j)^\phi, \tag{6}$$

where

$$M_\phi = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} (\mu_i \omega_j)^\phi. \tag{7}$$

From these expressions, the marginal probabilities of a form $p(s_i)$ and a counterpart $p(r_j)$ are obtained easily thanks to

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j)$$

$$p(r_j) = \sum_{i=1}^{n} p(s_i, r_j).$$

The cost of communication (Ω). The cost function is initially defined in Eq. 4 as in previous research (e.g. Ferrer-i-Cancho & Díaz-Guilera, 2007). In more detail,

$$\Omega = -\lambda I(S, R) + (1 - \lambda) H(S), \tag{8}$$

where $I(S, R)$ is the mutual information between forms from a repertoire S and counterparts from a repertoire R, and $H(S)$ is the entropy (or surprisal) of forms from a repertoire S. Knowing that $I(S, R) = H(S) + H(R) - H(S, R)$ (Cover & Thomas, 2006), the final expression for the cost function in this article is

$$\Omega(\lambda) = (1 - 2\lambda) H(S) - \lambda H(R) + \lambda H(S, R). \tag{9}$$

The entropies $H(S)$, $H(R)$ and $H(S, R)$ are easy to calculate applying the definitions of $p(s_i)$, $p(r_j)$ and $p(s_i, r_j)$, respectively.

The difference in the cost of learning a new word (Δ). There are two possible strategies to determine the counterpart with which a new form (a previously unlinked form) should connect (Fig. 3):

a. Connect the new form to a counterpart that is not already connected to any other forms.
b. Connect the new form to a counterpart that is connected to at least one other form.

The question we intend to answer is "when does strategy *a* result in a smaller cost than strategy *b*?" Or, in the terminology of child language research, "for which strategy is the assumption of mutual exclusivity more advantageous?" To answer these questions, we define Δ, as a the difference between the cost of each strategy. More precisely,

$$\Delta(\lambda) = \Omega'_a(\lambda) - \Omega'_b(\lambda), \tag{10}$$

where $\Omega'_a(\lambda)$ and $\Omega'_b(\lambda)$ are the new value of Ω when a new link is created using strategy *a* or *b* respectively. Then, our research question becomes "When is $\Delta < 0$?".

Formulae for $\Omega'_a(\lambda)$ and $\Omega'_b(\lambda)$ are derived in two steps. First, analyzing a general problem, i.e. Ω', the new value of Ω after producing a single mutation in A ("S1.2 Change in entropies after a single mutation in the adjacency matrix", Supplementary Information). Second, deriving expressions for the case where that mutation results from linking a new form (an unlinked form) to a counterpart, that can be linked or unlinked ("S1.3 Derivation of Δ", Supplementary Information).

Δ in two Classes of Skeleta

In previous work, the value of Δ was already calculated for $\phi = 0$, obtaining expressions equivalent to Eq. 5 (see "S1.3.1 The case $\varphi = 0$", Supplementary Information for a derivation). The next sections just summarize the more complex formulae that are obtained for each class of skeleta for $\phi \geq 0$ ("S1 The mathematical model in detail", Supplementary Information contains details on the derivation).

Vertex degrees do not exceed one

Here forms and counterparts both either have a single connection or are disconnected. Mathematically, this can be expressed as

$$\begin{aligned} \mu_i &\in \{0, 1\} \text{ for each } i \text{ such that } 1 \leq i \leq n \\ \omega_j &\in \{0, 1\} \text{ for each } j \text{ such that } 1 \leq j \leq m. \end{aligned}$$

Figure 3b offers a visual representation of a bipartite graph of this class. In case b, the counterpart we connect the new form to is connected to only one form ($\omega_j = 1$) and that form is connected to only one counterpart ($\mu_k = 1$). Under this class, Δ becomes

$$\Delta(\lambda) = (1 - 2\lambda)\left[\log\left(1 + \frac{2(2^\phi - 1)}{M + 1}\right) + \frac{2^{\phi+1}\log(2)\phi}{M + 2^{\phi+1} - 1}\right] - \lambda\frac{2^{\phi+1}\log(2)}{M + 2^{\phi+1} - 1}, \quad (11)$$

which can be rewritten as linear function of λ, i.e.

$$\Delta(\lambda) = a\lambda + b,$$

with

$$\begin{aligned} a &= 2\log\left(1 + \frac{2(2^\phi - 1)}{M + 1}\right) - (2\phi + 1)\frac{2^{\phi+1}\log(2)}{M + 2^{\phi+1} - 1} \\ b &= -\log\left(1 + \frac{2(2^\phi - 1)}{M + 1}\right) + \phi\frac{2^{\phi+1}\log(2)}{M + 2^{\phi+1} - 1}. \end{aligned}$$

Importantly, notice that this expression of Δ is determined only by λ, ϕ and M (the total number of links in the model). Thorough derivations can be found in "S1.3.3 Vertex degrees do not exceed one", Supplementary Information.

Counterpart degrees do not exceed one

This class of skeleta is a relaxation of the previous class. Counterparts are either connected to a single form or disconnected. Mathematically,

$$\omega_j \in \{0, 1\} \text{ for each } j \text{ such that } 1 \leq j \leq m.$$

Figure 3a offers a visual representation of a bipartite graph of this class. The number of forms the counterpart in case b is connected to is still 1 ($\omega_j = 1$) but this form may be connected to any number of counterparts; μ_k has to satisfy $1 \leq \mu_k \leq m$.

Under this class, Δ becomes

$$\begin{aligned}\Delta(\lambda) = (1-2\lambda)\Bigg\{ & \log\left(\frac{M_\phi + 1}{M_\phi + \left(2^\phi - 1\right)\mu_k^\phi + 2^\phi}\right) \\ & + \frac{1}{M_\phi + \left(2^\phi - 1\right)\mu_k^\phi + 2^\phi}\Bigg[(\phi+1)\frac{X(S,R)(2^\phi-1)(\mu_k^\phi+1)}{M_\phi+1} \\ & - \phi 2^\phi \log(2) + \mu_k^\phi\Big[\log(\mu_k)(\mu_k+\phi) \\ & - (\mu_k - 1 + 2^\phi)\log(\mu_k - 1 + 2^\phi)\Big]\Bigg]\Bigg\} \\ & - \frac{1}{M_\phi + \left(2^\phi - 1\right)\mu_k^\phi + 2^\phi}\Bigg[\lambda\left(\mu_k^\phi + 1\right)2^\phi \log\left(\mu_k^\phi + 1\right) \\ & - (1-\lambda)\phi 2^\phi \mu_k^\phi \log(\mu_k)\Bigg],\end{aligned} \tag{12}$$

where

$$X(S,R) = \sum_{i=1}^{n} \mu_i^{\phi+1} \log \mu_i \tag{13}$$

$$M_\phi = \sum_{i=1}^{n} \mu_i^{\phi+1}. \tag{14}$$

Equation 12 can also be expressed as a linear function of λ as

$$\Delta(\lambda) = a\lambda + b,$$

with

$$
\begin{aligned}
a \;=\; & 2\log\left(\frac{M_\phi + (2^\phi - 1)\mu_k^\phi + 2^\phi}{M_\phi + 1}\right) \\
& - \frac{1}{M_\phi + (2^\phi - 1)\mu_k^\phi + 2^\phi}\Big\{2^\phi\Big[(\mu_k^\phi + 1)\log(\mu_k^\phi + 1) + \phi\mu_k^\phi \log(\mu_k)\Big] \\
& + 2\Big[-(\phi+1)\frac{X(S,R)(2^\phi - 1)\mu_k^\phi + 1}{M_\phi + 1} \\
& + \phi 2^\phi \log(2) - \mu_k^\phi\big[\log(\mu_k)(\mu_k + \phi) - (\mu_k - 1 + 2^\phi)\log(\mu_k - 1 + 2^\phi)\big]\Big]\Big\} \\
b \;=\; & -\log\left(\frac{M_\phi + (2^\phi - 1)\mu_k^\phi + 2^\phi}{M_\phi + 1}\right) \\
& + \frac{1}{M_\phi + (2^\phi - 1)\mu_k^\phi + 2^\phi}\Big\{\phi 2^\phi \mu_k^\phi \log(\mu_k) - (\phi+1)\frac{X(S,R)(2^\phi - 1)\mu_k^\phi + 1}{M_\phi + 1} \\
& + \phi 2^\phi \log(2) - \mu_k^\phi\big[\log(\mu_k)(\mu_k + \phi) - (\mu_k - 1 + 2^\phi)\log(\mu_k - 1 + 2^\phi)\big]\Big\}.
\end{aligned}
$$

Being a relaxation of the previous class, the resulting expressions of Δ are more complex than those of the previous class, which are an in turn more complex than those of the case $\phi = 0$ Eq. 5. For further details on the derivation of Δ, see "S1.3.2 Counterpart degrees do not exceed one" in the Supplementary Information.

Notice that $X(S, R)$ Eq. 13 and M_ϕ Eq. 14 are determined by the degrees of the forms (μ_i's). To explore the phase space with a realistic distribution of μ_i's, we assume, without any loss of generality, that the μ_i's are sorted decreasingly, i.e. $\mu_1 \geq \mu_2 \geq ...\mu_i \geq \mu_{i+1} \geq ...\mu_n$. In addition, we assume

1. $\mu_n = 0$, because we are investigating the problem of linking and unlinked form with counterparts.
2. $\mu_{n-1} = 1$.
3. Form degrees are continuous.
4. The relationship between μ_i and its frequency rank is a right-truncated power-law, i.e.
$$\mu_i = ci^{-\tau} \tag{15}$$
for $1 \leq i \leq n-1$.

Section "S2 Form degrees and number of links" (Supplementary Information) shows that forms then follow Zipf's rank-frequency law, i.e.

$$p(s_i) = c' i^{-\alpha}$$

with

$$
\begin{aligned}
\alpha \;&=\; \tau(\phi + 1) \\
c' \;&=\; \frac{(n-1)^\alpha}{M_\phi}.
\end{aligned}
$$

The value of Δ is determined by λ, ϕ, μ_k and the sequence of degrees of the forms, which we have parameterized with α and n. When $\tau = \frac{\alpha}{\phi+1} = 0$, namely when

$\alpha = 0$ or when $\phi \to \infty$, we recover the class where vertex degrees do not exceed one but with just one form that is unlinked.

A continuous approximation to the number of edges gives ("S2 Form degrees and number of links", Supplementary Information)

$$M = (n-1)^{\frac{\alpha}{\phi+1}} \sum_{i=1}^{n-1} i^{-\frac{\alpha}{\phi+1}}. \tag{16}$$

We aim to shed some light on the possible trajectory that children will describe on Fig. S1 as they become older. One expects that M tends to increase as children become older, due to word learning. It is easy to see that Eq. 16 predicts that, if ϕ and α remain constant, M is expected to increase as n increases (Fig. S1). Besides, when n remains constant, a reduction of α implies a reduction of M when $\phi = 0$ but that effect vanishes for $\phi > 0$ (Fig. S1). Obviously, n tends to increase as a child becomes older (Saxton, 2010) and thus children's trajectory will be from left to right in Fig. S1. As for the temporal evolution of α, there are two possibilities. Zipf's pioneering investigations suggest that α remains close to 1 over time in English children (Zipf, 1949, Chapter IV). In contrast, a wider study reported a tendency of α to decrease over time in sufficiently old children of different languages (Baixeries et al., 2013) but the study did not determine the actual number of children where that trend was statistically significant after controlling for multiple comparisons. Then children, as they become older, are likely to move either from left to right, keeping α constant, or from the left-upper corner (high α, low n) to the bottom-right corner (low α, high n) within each panel of Fig. S1. When ϕ is sufficiently large, the actual evolution of some children (decrease of α jointly with an increase of n) is dominated by the increase of M that the growth of n implies in the long run (Fig. S1).

When exploring the space of parameters, we must warrant that μ_k does not exceed the maximum degree that n, ϕ and α yield, namely $\mu_k \leq \mu_1$, where μ_1 is defined according to Eq. 15 with $i = 1$, i.e.

$$\begin{aligned} \mu_k &\leq \mu_1 \\ &= c \\ &= (n-1)^{\tau} \\ &= (n-1)^{\frac{\alpha}{\phi+1}}. \end{aligned} \tag{17}$$

Results

Here we will analyze Δ, that takes a negative value when strategy a (linking a new form to a new counterpart) is more advantageous than strategy b linking a new form to an already connected counterpart), and a positive value otherwise. $|\Delta|$ indicates the strength of the bias towards strategy a if $\Delta < 0$; towards strategy b if $\Delta > 0$. Therefore, when $\Delta < 0$, the smaller the value of Δ, the higher the bias for strategy a whereas when $\Delta > 0$, the greater the value of Δ, the higher the bias for strategy b. Each class of skeleta is analyzed separately, beginning by the most restrictive class.

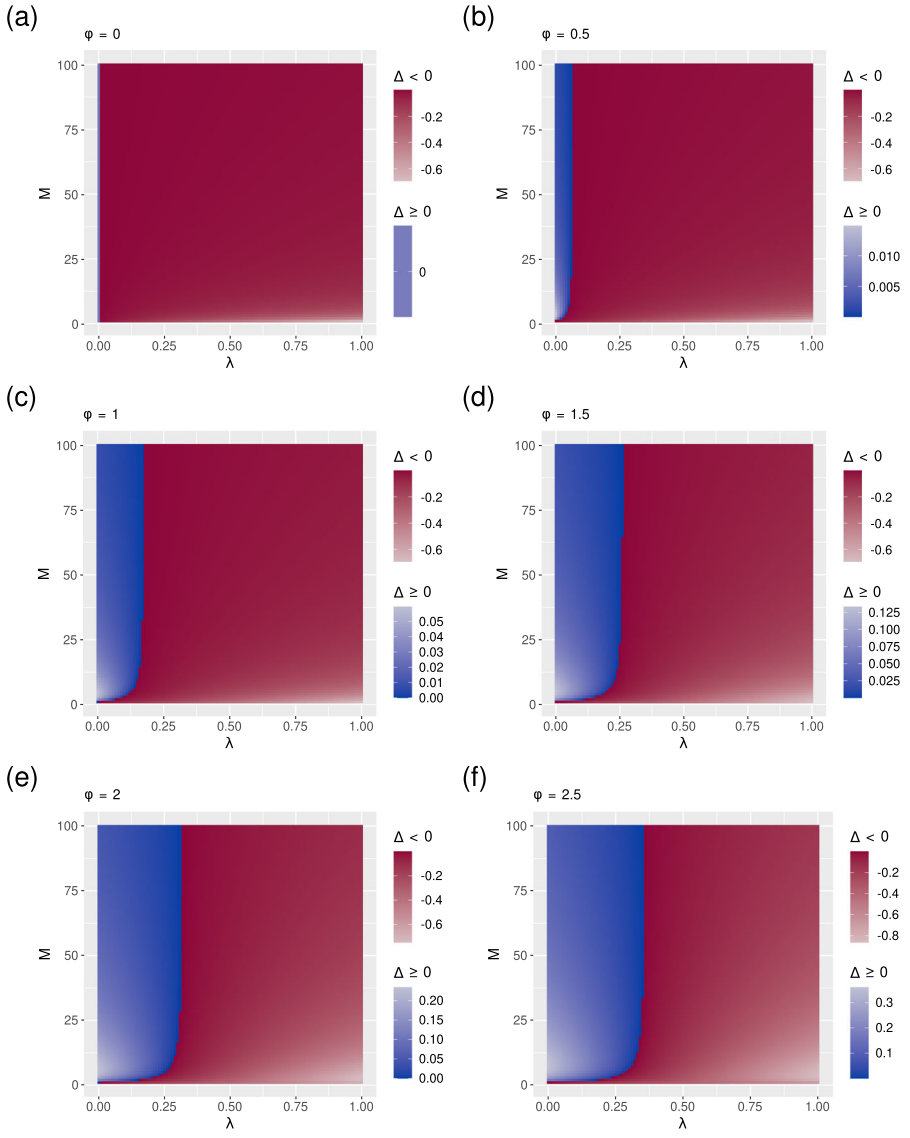

Fig. 4 Δ, the difference between the cost of strategy a and strategy b, as a function of M, the number of links and λ, the parameter that controls the balance between mutual information maximization and entropy minimization, when vertex degrees do not exceed one Eq. 11. Red indicates that strategy a is more advantageous while blue indicates that b is more advantageous. The lighter the red, the stronger the bias for strategy a. The lighter the blue, the stronger the bias for strategy b. **a** $\phi = 0$, **b** $\phi = 0.5$, **c** $\phi = 1$, **d** $\phi = 1.5$, **e** $\phi = 2$ and **f** $\phi = 2.5$

Vertex Degrees do not Exceed One

In this class of skeleta, corresponding to younger children, Δ depends only on ϕ, M and λ. We will explore the phase space with the help of two-dimensional heatmaps of Δ where the x-axis is always λ and the y-axis is M or ϕ.

Figure 4 reveals regions where strategy a is more advantageous (red) and regions where b is more advantageous (blue) according to Δ. The extreme situation is found when $\phi = 0$ where a single red region covers practically all space except for $\lambda = 0$ (Fig. 4, top-left) as expected from previous work (Ferrer-i-Cancho, 2017a) and Eq. 5.

Figure 5a summarizes these finding of regions, displaying the curve that defines the boundary between strategies a and b ($\Delta = 0$).

Figure 5b displays equivalent boundary curves summarizing Fig. S2, where ϕ replaces M on the y-axis of the heatmap. In Fig. 5b, each curve corresponds to a value of M and ϕ is placed on the y-axis.

Figure 5a and b show that strategy b is the optimal only if λ is sufficiently low, namely when the weight of entropy minimization is sufficiently high compared to that of mutual information maximization. Figure 5a shows that the larger the value of λ the larger the number of links (M) that is required for strategy b to be optimal. Figure 5a also indicates that the larger the value of ϕ, the broader the blue region where b is optimal. From a symmetric perspective, Fig. 5b shows that the larger the value of λ the larger the value of ϕ that is required for strategy b to be optimal and also that the larger the number of links (M), the broader the blue region where b is optimal.

Counterpart Degrees do not Exceed One

For this class of skeleta, corresponding to older children, we have assumed that word frequencies follow Zipf's rank-frequency law, namely the relationship between the probability of a form (the number of counterparts connected to each form) and its frequency rank follows a right-truncated power-law with exponent α ("The Mathematical Model"). Then Δ depends only on α (the exponent of the right-truncated power law), n (the number of forms), μ_k (the degree of the form linked to the counterpart in strategy b as shown in Fig. 3), ϕ and λ. We will explore the phase space with the help of two-dimensional heatmaps of Δ where the x-axis is always λ and the y-axis is μ_k, α or n. While in the class where vertex degrees do not exceed one we have found only one blue region (a region where $\Delta > 0$ meaning that b is more advantageous), this class yields up to two distinct blue regions located in opposite corners of the heatmap while keeping always a red region as shown in Fig. 6 for $\phi = 1$ from different perspectives. For the sake of brevity, this section only presents one set of heatmaps of Δ where $\phi = 1$ and μ_k varies on the y-axis (see "S3 Complementary figures" in the Supplementary Information for the remainder). A summary of the exploration of the parameter space follows.

Heatmaps of Δ as a function of λ and μ_k. The heatmaps of Δ for different combinations of parameters in Figs. 6, S3, S4, S5, S6 and S7 are summarized in Fig. 7, showing the frontiers between regions where $\Delta = 0$. Notice how, for $\phi = 0$, strategy

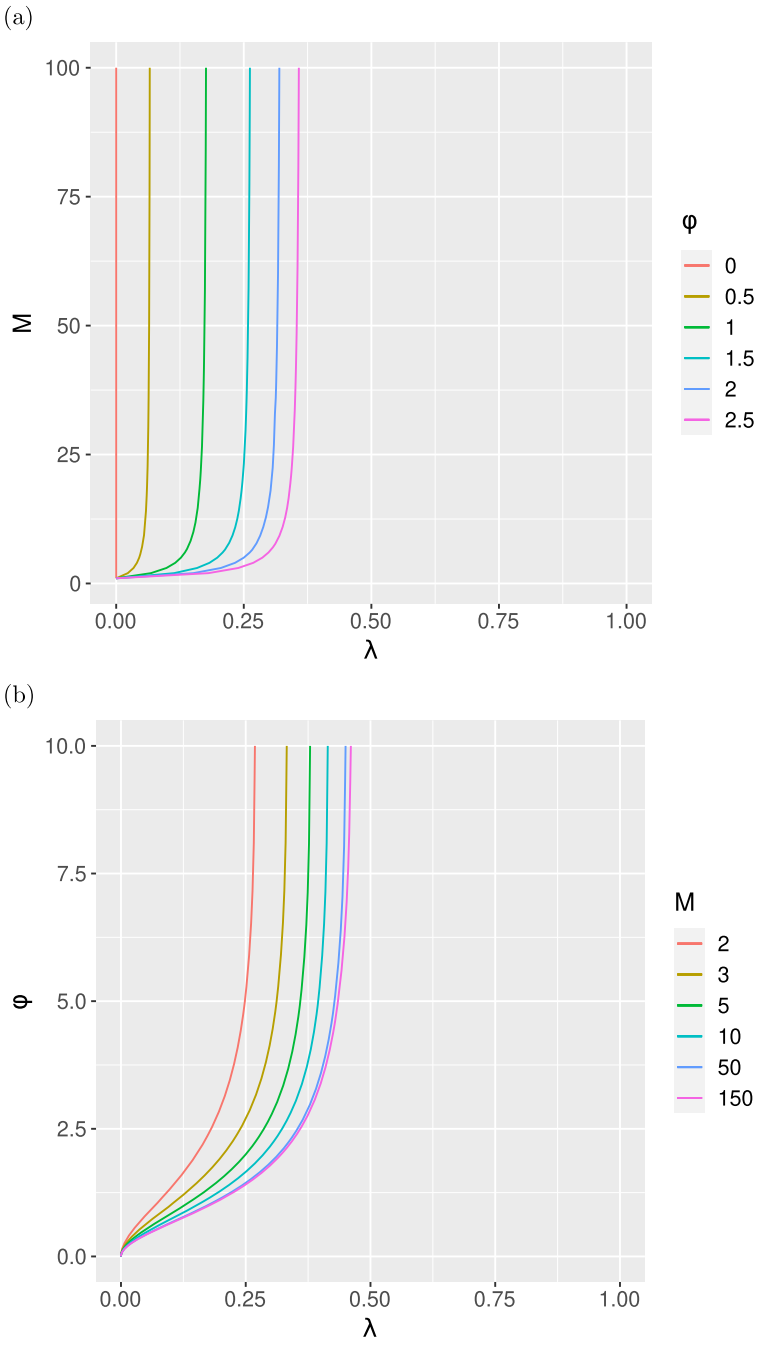

Fig. 5 Summary of the boundaries between positive and negative values of Δ when vertex degrees do not exceed one (Fig. 4 for **a**, Fig. S2 for **b**). Each curve shows the points where $\Delta = 0$ Eq. 12 as a function of λ for distinct values of ϕ. **a** has M on the y-axis and each curve belongs to a distinct value of ϕ, while **b** has ϕ on the y-axis and each curve belongs to a distinct value of M

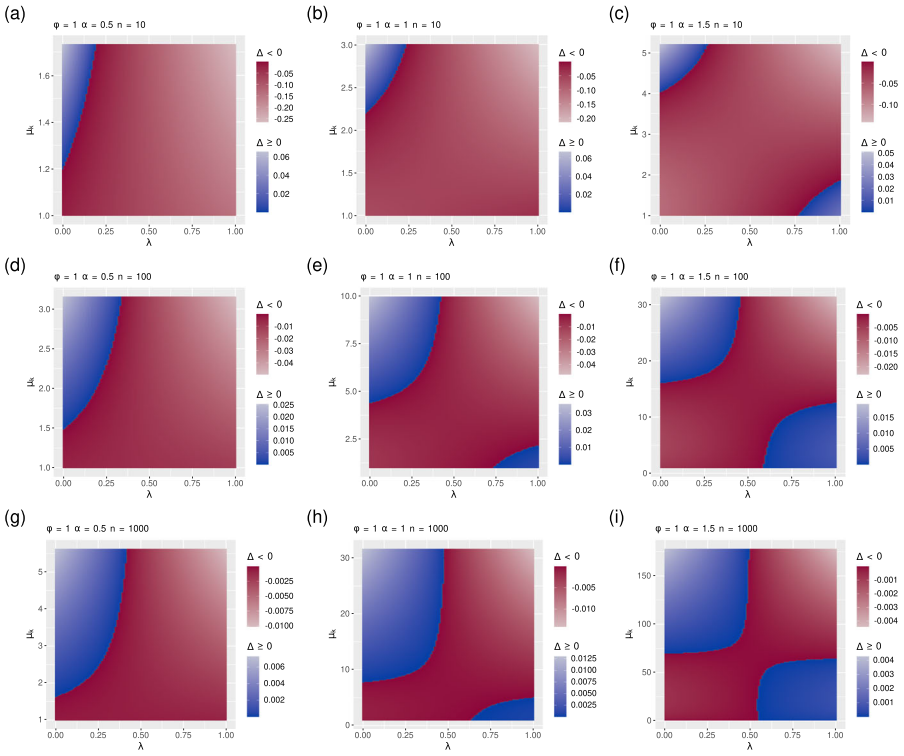

Fig. 6 Δ, the difference between the cost of strategy a and strategy b, as a function of μ_k, the degree of the form linked to the counterpart in strategy b as shown in Fig. 3, the number of links and λ, the parameter that controls the balance between mutual information maximization and entropy minimization, when the degrees of counterparts do not exceed one Eq. 11 and $\phi = 1$. Red indicates that strategy a is more advantageous while blue indicates that b is more advantageous. The lighter the red, the stronger the bias for strategy a. The lighter the blue, the stronger the bias for strategy b. Each heatmap corresponds to a distinct combination of n and α. The heatmaps are arranged, from left to right, with $\alpha = 0.5$, 1, 1.5 and, from top to bottom, with $n = 10$, 100, 1000. **a** $\alpha = 0.5$ and $n = 10$, **b** $\alpha = 1$ and $n = 10$, **c** $\alpha = 1.5$ and $n = 10$, **d** $\alpha = 0.5$ and $n = 100$, **e** $\alpha = 1$ and $n = 100$, **f** $\alpha = 1.5$ and $n = 100$, **g** $\alpha = 0.5$ and $n = 1000$, **h** $\alpha = 1$ and $n = 1000$, **i** $\alpha = 1.5$ and $n = 1000$

a is optimal for all values of $\lambda > 0$, as one would expect from Eq. 5. The remainder of the figures show how the shape of the two areas changes with each of the parameters. For small n and α, a single blue region indicates that strategy b is more advantageous than a when λ is closer to 0 and μ_k is higher. For higher n or α an additional blue region appears indicating that strategy b is also optimal for high values of λ and low values of μ_k.

Heatmaps of Δ as a function of λ and α. The heatmaps of Δ for different combinations of parameters in Figs. S8, S9, S10, S11 and S12 are summarized in Fig. S13, showing the frontiers between regions. There is a single region where strategy b is optimal for small values of μ_k and ϕ, but for larger values a second blue region appears.

Heatmaps of Δ as a function of λ and *n*. The heatmaps of Δ for different combinations of parameters in Figs. S14, S15, S16, S17 and S18 are summarized in Fig. S19. Again, one or two blue regions appear depending on the combination of parameters.

See "S4 Complementary figures with discrete degrees" in the Supplementary Information for the impact of using discrete form degrees on the results presented in this section.

Discussion

Vocabulary Learning

In previous research with $\phi = 0$, we predicted that the vocabulary learning bias (strategy a) would be present provided that mutual information minimization is not disabled ($\lambda > 0$) (Ferrer-i-Cancho, 2017a) as show in Eq. 5. However, the "decision" on whether assigning a new label to a linked or to an unlinked object is influenced by the age of a child and his/her degree of polylingualism. As for the effect of the latter, polylingual children tend to pick familiar objects more often than monolingual children, violating mutual exclusivity. This has been found for younger children below two years of age (17-22 months old in one study, 17-18 in another) (Houston-Price et al., 2010; Byers-Heinlein & Werker, 2013).

From three years onward, the difference between polylinguals and monolinguals either vanishes, namely both violate mutual exclusivity similarly (Nicoladis & Laurent, 2020; Frank & Poulin-Dubois, 2002), or polylingual children are still more willing to accept lexical overlap (Kalashnikova et al., 2015). One possible explanation for this phenomenon is the lexicon structure hypothesis (Byers-Heinlein & Werker, 2013), which suggests that children that already have many multiple-word-to-single-object mappings may be more willing to suspend mutual exclusivity.

As for the effect of age on monolingual children, the so-called mutual exclusivity bias has been shown to appear at an early age and, as time goes on, it is more easily suspended. Starting at 17 months old, children tend to look at a novel object rather than a familiar one when presented with a new word while 16-month-olds do not show a preference (Halberda, 2003). Interestingly, in the same study, 14-month-olds systematically look at a familiar object instead of a newer one. Reliance on mutual exclusivity is shown to improve between 18 and 30 months (Bion et al., 2013). Starting at least at 24 months of age, children may suspend mutual exclusivity to learn a second label for an object (Liittschwager & Markman, 1994). In a more recent study, it has been shown that three year old children will suspend mutual exclusivity if there are enough social cues present (Yildiz, 2020). Four to five year old children continue to apply mutual exclusivity to learn new words but are able to apply it flexibly, suspending it when given appropriate contextual information (Kalashnikova et al., 2016) in order to associate multiple labels to the same familiar object. As seen before, at 3 years of age both monolingual and polylingual children have similar willingness to suspend mutual exclusivity (Nicoladis & Laurent, 2020; Frank & Poulin-Dubois, 2002), although polylinguals may still have a greater tendency to accept multiple labels for the same object (Kalashnikova et al., 2015).

Here we have made an important contribution with respect to the precursor of the current model (Ferrer-i-Cancho, 2017a): we have shown that the bias is not theoretically inevitable (even when $\lambda > 0$) according a more realistic model. In a more complex setting, research on deep neural networks has shed light on the architectures, learning biases and pragmatic strategies that are required for the vocabulary learning bias to emerge (e.g. Gandhi & Lake, 2020; Gulordava et al., 2020). In "Results", we have discovered regions of the space of parameters where strategy a is not advantageous for two classes of skeleta. In the restrictive class, where one where vertex degrees do no exceed one, as expected in the earliest stages of vocabulary learning in children, we have unveiled the existence of a region of the phase space where strategy a is not advantageous (Figs. 5a and S2). In the broader class of skeleta where the degree of counterparts does not exceed one we have found up to two distinct regions where a is not advantageous (Figs. 7 and S13).

Crucially, our model predicts that the bias should be lost in older children. The argument is as follows. Suppose a child that has not learned a word yet. Then his skeleton belongs to the class where vertex degrees do not exceed one. Then suppose that the child learns a new word. It could be that he/she learns it following strategy a or b. If he applies b then the bias is gone at least for this word. Let us suppose that the child learns words adhering to strategy a for as long as possible. By doing this, he/she will increasing the number of links (M) of the skeleton keeping as invariant a one-to-one mapping between words and meanings (Figs. 1c and 2d), which satisfies that vertex degrees do not exceed one. Then Fig. 5a and b predict that the longer the time strategy a is kept (when $\phi > 0$) the larger the region of the phase space where a is not advantageous. Namely, as times goes on, it will become increasingly more difficult to keep a as the best option. Then it is not surprising that the bias weakens either in older children (e.g., Yildiz, 2020; Kalashnikova et al., 2016), as they are expected to have more links (larger M) because of their continued accretion of new words (Saxton, 2010), or in polylinguals (e.g., Nicoladis & Secco, 2000; Greene et al., 2013), where the mapping of words into meanings combining all their languages, is expected to yield more links than in monolinguals. Polylinguals make use of code-mixing to compensate for lexical gaps, as reported for from one-year-olds onward (Nicoladis & Secco, 2000) as well as in older children (five year olds) (Greene et al., 2013). As a result, the bipartite skeleton of a polylingual integrates the words and association in all the languages spoken and thus polylinguals are expected to have a larger value of M. Children who know more translation equivalents (words from different languages but with same meaning), adhere to mutual exclusivity less than other children (Byers-Heinlein & Werker, 2013). Therefore, our theoretical framework provides an explanation for the lexicon structure hypothesis (Byers-Heinlein & Werker, 2013), but shedding light on the possible origin of the mechanism, that is not the fact that there are already synonyms but rather the large number of links (Fig. 5b) as well as the capacity of words of higher degree to attract more meanings, a consequence of Eq. 3 with $\phi > 0$ in the vocabulary learning process (Fig. 3). Recall the stark contrast between Fig. 6 for $\phi = 1$ and Fig. S3 with $\phi = 0$, where such attraction effect is missing. Our models offer a transparent theoretical tool to understand the failure of deep neural networks to reproduce the vocabulary learning bias (Gandhi & Lake,

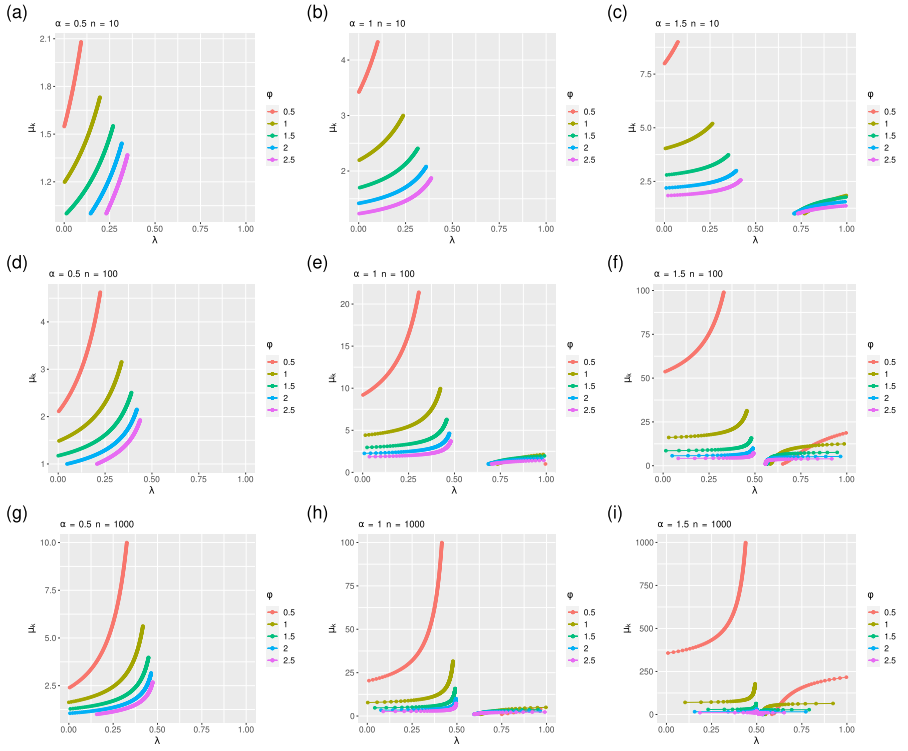

Fig. 7 Summary of the boundaries between positive and negative values of Δ when the degrees of counterparts do not exceed one (Figs. 6, S3, S4, S5, S6 and S7). Each curve shows the points where $\Delta = 0$ Eq. 13 as a function of λ and μ_k for distinct values of ϕ. **a** $\alpha = 0.5$ and $n = 10$, **b** $\alpha = 1$ and $n = 10$, **c** $\alpha = 1.5$ and $n = 10$, **d** $\alpha = 0.5$ and $n = 100$, **e** $\alpha = 1$ and $n = 100$, **f** $\alpha = 1.5$ and $n = 100$, **g** $\alpha = 0.5$ and $n = 1000$, **h** $\alpha = 1$ and $n = 1000$, **i** $\alpha = 1.5$ and $n = 1000$

2020): in its simpler form (vertex degrees do not exceed one), whether it is due to an excessive ϕ (Fig. 5a) or an excessive M (Fig. 5b).

We have focused on the loss of the bias in older children. However, there is evidence that the bias is missing initially in children, by the age of 14 months (Halberda, 2003). We speculate that this could be related to very young children having lower values of λ or larger values of ϕ as suggested by Figs. 5a and S2. This issue should be the subject of future research. Methods to estimate ϕ and λ in real speakers should be investigated.

Now we turn our attention to skeleta where only the degree of the counterparts does not exceed one, that we believe to be more appropriate for older children. Whereas ϕ, λ and M sufficed for the exploration of the phase space when vertex degrees do not exceed one, the exploration of that kind of skeleta involved many parameters: ϕ, λ, n, μ_k and α. The more general class exhibits behaviors that we have already seen in the more restrictive class. While an increase in M implies a widening of the region where a is not advantageous in the more restrictive class, the more general class experiences an increase of M when n is increased but α and ϕ

remain constant ("Counterpart degrees do not exceed one"). Consistently with the more restrictive class, such increase of M leads to a growth of the regions where a is not advantageous as it can be seen in Figs. 6, S4, S5, S6 and S7 when selecting a column (thus fixing α and ϕ) and moving from the top to the bottom increasing n. The challenge is that α may not remain constant in real children as they become older and how to involve the remainder of the parameters in the argument. In fact, some of these parameters are known to be correlated with child's age:

- n tends to increase over time in children, as children are learning new words over time (Saxton, 2010). We assume that the loss of words can be neglected in children.
- M tends to increase over time in children. In this class of skeleta, the growth of M has two sources: the learning of new words as well as the learning of new meanings for existing words. We assume that the loss of connections can be neglected in children.
- The ambiguity of the words that children learn over time tends to increase over time (Casas et al., 2018). This does not imply that children are learning all the meanings of the word according to some online dictionary but rather than as times go on, children are able to handle words that have more meanings according to adult standards.
- α remains stable over time or tends to decrease over time in children depending on the individual (Baixeries et al., 2013; Zipf, 1949, Chapter IV).

For other parameters, we can just speculate on their evolution with child's age. The growth of M and the increase in the learning of ambiguous words over time leads to expect that the maximum value of μ_k will be larger in older children. It is hard to tell if older children will have a chance to encounter larger values of μ_k. We do not know the value of λ in real language but the higher diversity of vocabulary in older children and adults (Baixeries et al., 2013) suggests that λ may tend to increase over time, because the lower the value of λ, the higher the pressure to minimize the entropy of words Eq. 4, namely the higher the force towards unification in Zipf's view (Zipf, 1949). We do not know the real value of ϕ but a reasonable choice for adult language is $\phi = 1$ (Ferrer-i-Cancho & Vitevitch, 2018).

Given the complexity of the space of parameters in the more general class of skeleta where only the degrees of counterparts cannot exceed one, we cannot make predictions that are as strong as those stemming from the class where vertex degrees cannot exceed one. However, we wish to make some remarks suggesting that a weakening of the vocabulary learning bias is also expected in older children for this class (provided that $\phi > 0$). The combination of increasing n and a value of α that is stable over time suggests a weakening of the strategy a over time from different perspectives

- Children evolve on a column of panels (constant α) of the matrix of panels in Figs. 6, S4, S5, S6 and S7, moving from top (low n) to the bottom (large n). That trajectory implies an increase of the size of the blue region, where strategy a is not advantageous.
- We do not know the temporal evolution of μ_k but once μ_k is fixed, namely a row of panels is selected in Figs. S8, S9, S10, S11 and S12, children evolve from left

(lower n) to right (higher n), which implies an increase of the size of the blue region where strategy a is not advantageous as children become older.
– Within each panel in Figs. S14, S15, S16, S17 and S18, an increase of n, as a results of vocabulary learning over time, implies a widening of the blue region.

In the preceding analysis we have assumed that α remains stable over time. We wish to speculate on the combination of increasing n and decreasing α as time goes on in certain children. In that case, children would evolve close to the diagonal of the matrix of panels, starting from the right-upper corner (low n, high α, panel (c)) towards the lower-left corner (high n, low α, panel (g)) in Figs. 6, S4, S5, S6 and S7, which implies an increase of the size of the blue region where strategy a is not advantageous. Recall that we have argued that a combined increase of n and decrease of α is likely to lead in the long run to an increase of M (Fig. S1). We suggest that the behavior "along the diagonal" of the matrix is an extension of the weakening of the bias when M is increased in the more restrictive class (Fig. 5b).

In our exploration of the phase space for the class of the skeleta where the degrees of counterparts do not exceed one, we assumed a right-truncated power-law with two parameters, α and n as a model for Zipf's rank-frequency law. However, distributions giving a better fit have been considered (Li et al., 2010) and function (distribution) capturing the shape of the law of what Piotrowski called saturated samples (Piotrowski & Spivak, 2007) should be considered in future research. Our exploration of the phase space was limited by a brute force approach neglecting the negative correlation between n and α that is expected in children where α and time are negatively correlated: as children become older, n increases as a result of word learning (Saxton, 2010) but α decreases (Baixeries et al., 2013). A more powerful exploration of the phase space could be performed with a realistic mathematical relationship of the expected correlation between n and α, which invites to empirical research. Finally, there might be deeper and better ways of parameterizing the class of skeleta.

Biosemiotics

Biosemiotics is concerned about building bridges between biology, philosophy, linguistics, and the communication sciences as announced in the front page of this journal. As far as we know, there is little research on the vocabulary learning bias in other species. Its confirmation in a domestic dog suggests that "*the perceptual and cognitive mechanisms that may mediate the comprehension of speech were already in place before early humans began to talk*" (Kaminski et al., 2004). We hypothesize that the cost function Ω captures the essence of these mechanisms. A promising target for future research are ape gestures, where there has been significant progress recently on their meaning (Hobaiter & Byrne, 2014). As far as we know, there is no research on that bias in other domains that also fall into the scope of biosemiotics, e.g., in unicellular organisms such as bacteria. Our research has established some mathematical foundations for research on the accretion and interpretation of signs across the living world, not only among great apes, a key problem in research program of biosemiotics (Kull, 2018).

The remainder of the discussion section is devoted to examine general challenges that are shared by biosemiotics and quantitative linguistics, a field that, as biosemiotics, aspires to contribute to develop a science beyond human communication.

Science and its Method

It has been argued that a problem of research on the rank-frequency is law is the *The absence of novel predictions...* which *has led to a very peculiar situation in the cognitive sciences, where we have a profusion of theories to explain an empirical phenomenon, yet very little attempt to distinguish those theories using scientific methods.* (Piantadosi, 2014). As we have already shown the predictive power of a model whose original target was the rank-frequency laws here and in previous research (Ferrer-i-Cancho, 2017a), we take this criticism as an invitation to reflect on science and its method (Altmann, 1993; Bunge, 2001).

The generality of the patterns for theory construction

While in psycholinguisics and the cognitive sciences a major source of evidence are often experiments involving restricted tasks or sophisticated statistical analyses covering a handful of languages (typically English and a few other Indo-European languages), quantitative linguistics aims to build theory departing from statistical laws holding in a typologically wide range of languages (Köhler, 1987; Debowski, 2020) as reflected in Fig. 1. In addition, here we have investigated a specific vocabulary learning phenomenon that is, however, supported cross-linguistically (recall "Introduction"). A recent review on the efficiency of languages, only pays attention to the law of abbreviation (Gibson et al., 2019) in contrast with the body of work that has been developed in the last decades linking laws with optimization principles (Fig. 1), suggesting that this law is the only general pattern of languages that is shaped by efficiency or that linguistic laws are secondary for deep theorizing on efficiency. In other domains of the cognitive sciences, the importance of scaling laws has been recognized (Chater & Brown, 1999; Kello et al., 2010; Baronchelli et al., 2013).

Novel predictions

In "Vocabulary Learning", we have checked predictions of our information theoretic framework that matches knowledge on the vocabulary learning bias from past research. Our theoretical framework allows the researcher to play the game of science in another direction: use the relevant parameters to guide the design of new experiments with children or adults where more detailed predictions of the theoretical framework can be tested. For children who have about the same n and α, and $\phi = 1$, our model predicts that strategy a will be discarded if (Fig. 6)

(1) λ is low and μ_k (Fig.3) is large enough.
(2) λ is high and μ_k is sufficiently low.

Interestingly, there is a red horizontal band in Fig. 6, and even for other values of ϕ such that $\phi \neq 1$ but keeping $\phi > 0$ (Figs. S4, S5, S6, S7), indicating the existence of some value of μ_k or a range of μ_k where strategy a is always advantageous (notice however, that when $\phi > 1$, the band may become too narrow for an integer μ_k to fit as suggested by Figs. S23, S24, S25 in the Supplementary Information, Section "S4 Complementary figures with discrete degrees"). Therefore the 1st concrete prediction is that, for a given child, there is likely to be some range or value of μ_k where the bias (strategy a) will be observed. The 2nd concrete prediction that can be made is on the conditions where the bias will not be observed. Although the true value of λ is not known yet, previous theoretical research with $\phi = 0$ suggests that $\lambda \leq 1/2$ in real language (Ferrer-i-Cancho & Sole, 2003; Ferrer-i-Cancho, 2005b; 2006; 2005a), which would imply that real speakers should satisfy only (1). Child or adult language researchers may design experiments where μ_k is varied. If successful, that would confirm the lexicon structure hypothesis (Byers-Heinlein & Werker, 2013) but providing a deeper understanding. These are just examples of experiments that could be carried out.

Towards a mathematical theory of language efficiency

Our past and current research on the efficiency are supported by a cost function and a (analytical or numerical) mathematical procedure that links the minimization of the cost function with the target phenomena, e.g., vocabulary learning, as in research on how pressure for efficiency gives rise to Zipf's rank-frequency law, the law of abbreviation or Menzerath's law (Gustison et al., 2016; Ferrer-i-Cancho, 2005b, 2019). In the cognitive sciences, such a cost function and the mathematical linking argument are sometimes missing (e.g., Piantadosi et al., 2011) and neglected when reviewing how languages are shaped by efficiency (Gibson et al., 2019). A truly quantitative approach in the context of language efficiency is two-fold: it has to comprise either a quantitative description of the data and a quantitative theorizing, i.e. it has to employ both statistical methods of analysis and mathematical methods to define the cost and the how cost minimization leads to the expected phenomena. Our framework relies on standard information theory (Cover & Thomas, 2006) and its extensions (Ferrer-i-Cancho et al., 2019; Debowski, 2020). The psychological foundations of the information theoretic principles postulated in that framework and the relationships between them have already been reviewed (Ferrer-i-Cancho, 2018). How the so-called noisy-channel "theory" or noisy-channel hypothesis explains the results in (Piantadosi et al., 2011), others reviewed recently (Gibson et al., 2019) or language laws in a broad sense has not yet shown, to our knowledge, with detailed enough information theory arguments. Furthermore, the major conclusions of the statistical analysis of Piantadosi et al. (2011) have recently been shown to change substantially after improving the methods: effects attributable to plain compression are stronger than previously reported (Meylan & Griffiths, 2021). Theory is crucial to reduce false positives and replication failures (Stewart & Plotkin, 2021). In addition, higher order compression can explain more parsimoniously phenomena that are central in noisy-channel "theorizing" (Ferrer-i-Cancho, 2017b).

The trade-off between parsimony and perfect fit

Our emphasis is on generality and parsimony over perfect fit. Piantadosi (2014) makes emphasis on what models of Zipf's rank-frequency law apparently do not explain while our emphasis is on what the models do explain and the many predictions they make (Table 1), in spite of their simple design. It is worth reminding a big lesson from machine learning, i.e. a perfect fit can be obtained simply by overfitting the data and another big lesson from the philosophy of science to machine learning and AI: sophisticated models (specially deep learning ones) are in most cases black boxes that imitate complex behavior but neither explain nor yield understanding. In our theoretical framework, the principle of contrast (Clark, 1987) or the mutual exclusivity bias (Markman & Wachtel, 1988; Merriman & Bowman, 1989) are not principles *per se* (or core principles) but predictions of the principle of mutual information maximization involved in explaining the emergence of Zipf's rank-frequency law (Ferrer-Cancho 2003, 2005b) and word order patterns (Ferrer-i-Cancho, 2017b). Although there are computational models that are able to account for that vocabulary learning bias and other phenomena (Frank et al., 2009; Gulordava et al., 2020), ours is much simpler, transparent (in opposition to black box modeling) and to the best our knowledge, the first to predict that the bias will weaken over time providing a preliminary understanding of why this could happen.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12304-021-09452-w.

Acknowledgements We are grateful to two anonymous reviewers for their valuable feeback and recommendations to improve the article. We are also grateful to A. Hernández-Fernández and G. Boleda for their revision of the article and many recommendations to improve it. The article has benefited from discussions with T. Brochhagen, S. Semple and M. Gustison. Finally, we thank C. Hobaiter for her advice and inspiration for future research. DCC and RFC are supported by the grant TIN2017-89244-R from MINECO (Ministerio de Economía, Industria y Competitividad). RFC is also supported by the recognition 2017SGR-856 (MACDA) from AGAUR (Generalitat de Catalunya).

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Altmann, G. (1993). Science and linguistics. In R. Röhler, & B. Rieger (Eds.) *Linguistics, Contributions to Quantitative* (pp. 3–10). Dordrecht: Kluwer.

Baixeries, J., Elvevåg, B., & Ferrer-i-Cancho, R. (2013). The evolution of the exponent of Zipf's law in language ontogeny. *PLoS One*, *8*(3), e53227.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chatter, N., & Christiansen, M. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, *17*, 348–360.

Bentz, C., & Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In C. Bentz, G. Jäger, & I. Yanovich (Eds.) *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tübingen.

Bion, R. A., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word-object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, *126*(1), 39–53. https://doi.org/10.1016/j.cognition.2012.08.008.

Brochhagen, T. (2021). Brief at the risk of being misunderstood: Consolidating population- and individual-level tendencies. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-021-00099-x.

Bunge, M. (2001). *La science, sa méthode et sa philosophie*: Vigdor.

Byers-Heinlein, K., & Werker, J. F. (2013). Lexicon structure and the disambiguation of novel words: Evidence from bilingual infants. *Cognition*, *128*(3), 407–416. https://doi.org/10.1016/j.cognition.2013.05.010.

Casas, B., Català, N., Ferrer-i-Cancho, R., Hernández-fernández, A., & Baixeries, J. (2018). The polysemy of the words that children learn over time. *Interaction Studies*, *19*(3), 389–426.

Chater, N., & Brown, G. D. A. (1999). Scale invariance as a unifying psychological principle. *Cognition*, *69*, 1999.

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.) *Mechanisms of language acquisition*. Hillsdale: Lawrence Erlbaum Associates.

Clark, E. (1993). *The lexicon in acquisition*: Cambridge University Press.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*, 2nd edn. New York: Wiley.

Dangli, L., & Abazaj, G. (2009). Absolute versus relative synonymy. *Linguistic and Communicative Performance Journal*, *2*, 64–68.

Deacon, T. W. (1997). *The Symbolic Species: the Co-evolution of Language and the Brain. W W*. New York: Norton & Company.

Deacon, T. W. (2015). Steps to a science of biosemiotics. *Green Letters*, *19*(3), 293–311. https://doi.org/10.1080/14688417.2015.1072948.

Debowski, L. (2020). *Information theory meets power laws: Stochastic processes and language models*. Hoboken: Wiley.

Eco, U. (1986). *Semiotics and the philosophy of language*. Bloomington: Indiana University Press.

Ellis, S. R., & Hitchcock, R. J. (1986). The emergence of Zipf's law: spontaneous encoding by users of a command language. *IEEE Trans Syst Man Cyber*, *16*(3), 423–427.

Eun-Nam, S. (2017). Word learning characteristics of 3-to 6-year-olds: Focused on the mutual exclusivity assumption. *Journal of Speech-Language & Hearing Disorders*, *26*(4), 33–40.

Ferrer-i-Cancho, R. (2005a). The variation of Zipf's law in human language. *European Physical Journal B*, *44*, 249–257.

Ferrer-i-Cancho, R. (2005b). Zipf's law from a communicative phase transition. *European Physical Journal B*, *47*, 449–457. https://doi.org/10.1140/epjb/e2005-00340-y.

Ferrer-i-Cancho, R. (2006). When language breaks into pieces. A conflict between communication through isolated signals and language. *Bio Systems*, *84*, 242–253.

Ferrer-i-Cancho, R. (2016a). Compression and the origins of Zipf's law for word frequencies. *Complexity*, *21*, 409–411.

Ferrer-i-Cancho, R. (2016b). The meaning-frequency law in Zipfian optimization models of communication. *Glottometrics*, *35*, 28–37.

Ferrer-i-Cancho, R. (2017a). The optimality of attaching unlinked labels to unlinked meanings. *Glottometrics*, *36*, 1–16.

Ferrer-i-Cancho, R. (2017b). The placement of the head that maximizes predictability. An information theoretic approach. *Glottometrics*, *39*, 38–71.

Ferrer-i-Cancho, R. (2018). Optimization models of natural communication. *Journal of Quantitative Linguistics*, *25*(3), 207–237.

Ferrer-i-Cancho, R., & Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment, 06009*(6), https://doi.org/10.1088/1742-5468/2007/06/P06009.

Ferrer-i-Cancho, R., & Sole, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America, 100*(3):788–791, https://doi.org/10.1073/pnas.0335980100.

Ferrer-i-Cancho, R., & Vitevitch, M. (2018). The origins of Zipf's meaning-frequency law. *Journal of the American Association for Information Science and Technology*, *69*(11), 1369–1379.
Ferrer-i-Cancho, R., Riordan, O., & Bollobás, B. (2005). The consequences of Zipf's law for syntax and symbolic reference. *Proceedings of the Royal Society of London B*, *272*, 561–565.
Ferrer-i-Cancho, R., Bentz, C., & Seguin, C. (2019). Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics in press*. https://doi.org/10.1080/09296174.2020.1778387.
Frank, I., & Poulin-Dubois, D. (2002). Young monolingual and bilingual children's responses to violation of the mutual exclusivity principle. *International Journal of Bilingualism*, *6*(2), 125–146. https://doi.org/10.1177/13670069020060020201.
Frank, M. C., Goodman, N. D., & Tenenbaum, J.B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585. https://doi.org/10.1111/j.1467-9280.2009.02335.x.
Fromkin, V., Rodman, R., & Hyams, N. (2014). *An introduction to language*, 1st edn. Boston: Wadsworth Publishing.
Futrell, R. (2020). https://twitter.com/rljfutrell/status/1275834876055351297.
Gandhi, K., & Lake, B. (2020). Mutual exclusivity as a challenge for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 33.
Genty, E., & Zuberbühler, K. (2014). Spatial reference in a bonobo gesture. *Current Biology*, *24*(14), 1601–1605. https://doi.org/10.1016/j.cub.2014.05.065.
Gibson, E., Futrell, R., Piantadosi, S., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*, 389–407.
Greene, K. J., Peña, E. D., & Bedore, L.M. (2013). Lexical choice and language selection in bilingual preschoolers. *Child Language Teaching and Therapy*, *29*(1), 27–39. https://doi.org/10.1177/0265659012459743.
Gulordava, K., Brochhagen, T., & Boleda, G. (2020). Deep daxes: Mutual exclusivity arises through both learning biases and pragmatic strategies in neural networks. In *Proceedings of CogSci*, (Vol. 2020 pp. 2089–2095).
Gustison, M. L., Semple, S., Ferrer-i-Cancho, R., & Bergman, T. (2016). Gelada vocal sequences follow Menzerath,'s linguistic law. *Proceedings of the National Academy of Sciences USA*, *13*(19), E2750–E2758. https://doi.org/10.1073/pnas.1522072113.
Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), 23–34. https://doi.org/10.1016/S0010-0277(02)00186-5.
Haryu, E. (1991). A developmental study of children's use of mutual exclusivity and context to interpret novel words. *The Japanese Journal of Educational Psychology, 39*(1):11–20, https://doi.org/10.5926/jjep1953.39.1_11.
Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. *Cognition*, *189*(February), 11–22. https://doi.org/10.1016/j.cognition.2019.03.005.
Hobaiter, C., & Byrne, R. W. (2014). The meanings of chimpanzee gestures. *Current Biology*, *24*, 1596–1600.
Houston-Price, C., Caloghiris, Z., & Raviglione, E. (2010). Language experience shapes the development of the mutual exclusivity bias. *Infancy*, *15*(2), 125–150. https://doi.org/10.1111/j.1532-7078.2009.00009.x.
Hung, W. Y., Patrycia, F., & Yow, W.Q. (2015). Bilingual children weigh speaker's referential cues and word-learning heuristics differently in different language contexts when interpreting a speaker's intent. *Frontiers in Psychology*, *6*(JUN), 1–9. https://doi.org/10.3389/fpsyg.2015.00796.
Hurford, J. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua, 77*:187–222. https://doi.org/10.1016/0024-3481(89)90015-6.
Kalashnikova, M., Mattock, K., & Monaghan, P. (2015). The effects of linguistic experience on the flexible use of mutual exclusivity in word learning. *Bilingualism*, *18*(4), 626–638. https://doi.org/10.1017/S1366728914000364.
Kalashnikova, M., Mattock, K., & Monaghan, P. (2016). Flexible use of mutual exclusivity in word learning. *Language Learning and Development*, *12*(1), 79–91. https://doi.org/10.1080/15475441.2015.1023443.
Kalashnikova, M., Oliveri, A., & Mattock, K. (2019). Acceptance of lexical overlap by monolingual and bilingual toddlers. *International Journal of Bilingualism*, *23*(6), 1517–1530. https://doi.org/10.1177/1367006918808041.
Kaminski, J., Call, J., & Fischer, J. (2004). Word learning in a domestic dog: Evidence for "fast mapping". *Science*, *304*(5677), 1682–1683. https://doi.org/10.1126/science.1097859.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.
Kello, C. T., Brown, G. D. A., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Orden, G.C.V. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, *14*(5), 223–232. https://doi.org/10.1016/j.tics.2010.02.005.
Köhler, R. (1987). System theoretical linguistics. *Theor Linguist*, *14*(2-3), 241–257.
Kull, K. (1999). Biosemiotics in the twentieth century: A view from biology. *Semiotica*, *127*(1/4), 385–414.
Kull, K. (2018). Choosing and learning: Semiosis means choice. *Sign Systems Studies*, *46*(4), 452–466.
Kull, K. (2020). Codes: Necessary, but not sufficient for meaning-making. *Constructivist Foundations*, *15*(2), 137–139.
Li, W., Miramontes, P., & Cocho, G. (2010). Fitting ranked linguistic data with two-parameter functions. *Entropy*, *12*(7), 1743–1764.
Liittschwager, J. C., & Markman, E. M. (1994). Sixteen-and 24-month-olds' use of mutual exclusivity as a default assumption in second-label learning. *Developmental Psychology*, *30*(6), 955–968. https://doi.org/10.1037/0012-1649.30.6.955.
Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods. *Instruments, and Computers*, *28*(2), 203–208.
Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.
Merriman, W. W., & Bowman, L. L. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development*, *54*, 1–129.
Meylan, S., & Griffiths, T. (2021). The challenges of large-scale, web-based language datasets: Word length and predictability revisited. PsyArXiv https://doi.org/10.31234/osf.io/6832r, psyarxiv.com/6832r.
Moore, R. (2014). Ape gestures: Interpreting chimpanzee and bonobo minds. *Current Biology*, *24*(14), R645–R647. https://doi.org/10.1016/j.cub.2014.05.072.
Nicoladis, E., & Laurent, A. (2020). When knowing only one word for "car" leads to weak application of mutual exclusivity. *Cognition 196*, *104087*, 2019. https://doi.org/10.1016/j.cognition.2019.104087.
Nicoladis, E., & Secco, G. (2000). The role of a child's productive vocabulary in the language choice of a bilingual family. *First Language*, *20*(58), 003–28. https://doi.org/10.1177/014272370002005801.
Piantadosi, S. (2014). Zipf's law in natural language: a critical review and future directions. *Psychonomic Bulletin and Review*, *21*, 1112–1130.
Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.
Piotrowski, R. G., & Spivak, D. L. (2007). Linguistic disorders and pathologies: synergetic aspects. In P. Grzybek, & R. Köhler (Eds.) *Exact methods in the study of language and text. to honor gabriel altmann* (pp. 545–554). Berlin: Gruyter.
Pulvermüller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, *17*(9), 458–470. https://doi.org/10.1016/j.tics.2013.06.004.
Pulvermüller, F. (2001). Brain reflections of words and their meaning. *Trends in Cognitive Sciences*, *5*(12), 517–524.
Saxton, M. (2010). *Child language. Acquisition and development*. Los Angeles: SAGE. Chap 6. The developing lexicon: what's in a name? pp. 133–158.
Steels, L. (1996). The spontaneous self-organization of an adaptive language. *Machine Intelligence*, *15*, 205–224.
Stewart, A. J., & Plotkin, J. B. (2021). The natural selection of good science. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01111-x.
Yildiz, M. (2020). Conflicting nature of social-pragmatic cues with mutual exclusivity regarding three-year-olds' label-referent mappings. *Psychology of Language and Communication*, *24*(1), 124–141. https://doi.org/10.2478/plc-2020-0008.
Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in crosssituational statistical learning. *Proceedings of the annual meeting of the cognitive science society*, 715–720.
Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942. https://doi.org/10.1073/pnas.1800521115.

Zaslavsky, N., Maldonado, M., & Culbertson, J. (2021). Let's talk (efficiently) about us: Person systems achieve near-optimal compression. PsyArXiv https://doi.org/10.31234/osf.io/kcu27, psyarxiv.com/kcu27.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology*, *33*, 251–266.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge: Addison-Wesley.