Available online at www.sciencedirect.com
SCIENCE DIRECT®
Artificial Intelligence 167 (2005) 206–242
Artificial Intelligence
www.elsevier.com/locate/artint

# The emergence of compositional structures in perceptually grounded language games

Paul Vogt [a,b,1]

[a] *Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, UK*
[b] *Induction of Linguistic Knowledge group/Computational Linguistics, Tilburg University, The Netherlands*

Received 28 July 2004; received in revised form 20 February 2005; accepted 14 April 2005

Available online 19 August 2005

**Abstract**

This paper describes a new model on the evolution and induction of compositional structures in the language of a population of (simulated) robotic agents. The model is based on recent work in language evolution modelling, including the iterated learning model, the language game model and the Talking Heads experiment. It further adopts techniques recently developed in the field of grammar induction. The paper reports on a number of different experiments done with this new model and shows certain conditions under which compositional structures can emerge. The paper confirms previous findings that a transmission bottleneck serves as a pressure mechanism for the emergence of compositionality, and that a communication strategy for guessing the references of utterances aids in the development of qualitatively 'good' languages. In addition, the results show that the emerging languages reflect the structure of the world to a large extent and that the development of a semantics, together with a competitive selection mechanism, produces a faster emergence of compositionality than a predefined semantics without such a selection mechanism.


*E-mail address:* paulv@ling.ed.ac.uk (P. Vogt).

[1] This work has been sponsored by the European Commission through a Marie Curie Fellowship and by the Netherlands Organisation for Scientific Research through a Veni grant. The author wishes to thank Simon Kirby, Ehud Reiter, Andrew D.M. Smith, Kenny Smith and three anonymous reviewers for their invaluable comments made on earlier versions of this manuscript.

0004-3702/$ – see front matter 
doi:10.1016/j.artint.2005.04.010

*Keywords:* Compositionality; Grammar induction; Grounding; Iterated learning; Language evolution; Language games

## 1. Introduction

One recent trend in models of language learning is the emergence of an increasing number of models simulating certain aspects of the origins and evolution of language, see [9,11,25,45] for overviews. This paper presents a new model to study the emergence and dynamics of compositional structures in languages whose semantics are connected with a simulated world. Although this world is far from realistic, it can be—and has been—implemented physically using real robots [49].

The ability to form compositional structures—as part of syntax—is one of the key aspects of human language. Here is a definition of compositionality that was taken from a web-site of a recent series of conferences on compositionality:[2] "Compositionality is a key feature of structured representational systems, be they linguistic, mental or neuronal. A system of representations is compositional just in case the semantic values of complex representations are determined by the semantic values of their parts." For instance, the sentence "give me the book" can be described semantically as `give(me,thebook)`, where the word "give" maps onto the action `give`, "me" onto the person `me` and "the book" onto the object `thebook`. In contrast, *holistic* expressions have no structural relations between parts of the expressions and parts of their meanings. In "kicked the bucket" when used to mean `died`, for instance, no part of the expression has a relation to any part of the meaning `died` (apart from the aspect of tense).[3]

One of the frequently asked questions in studies on language origins and evolution is: how could compositional structures in human languages have emerged? One line of research assumes that compositional structures emerged from exploiting (e.g. random) regularities found in protolanguages based on holophrases [67], a line of research that has been adopted by many computational modellers, most notably [6,20,23]. In these studies, it was shown that compositional structures in language can emerge when the learning examples do not cover the entire language (i.e., there was a *bottleneck* on the transmission of language [20]), provided the learners have a predefined mechanism for acquiring compositional structures. Other researchers have assumed that the ability to use syntax has evolved as a biological adaptation [34], as modelled in, e.g., [8]. Yet other modellers have assumed that compositional structures can emerge based on competition between exemplars [2] and self-organisation in a production system [46]. Note that in all computer models developed so far, learning mechanisms have been implemented that can acquire compositional structures. Hence, all studies use the assumption that a specialised learning mechanism has evolved prior to the 'emergence' of compositional languages, and therefore investigate the

[2] See http://www.phil-fak.uni-duesseldorf.de/thphil/compositionality/ and http://www.cognition.ens.fr/nac2004/.

[3] Note that these examples are extreme and "easy" examples of compositionality. There are many other interesting aspects of compositionality that are not covered by these examples. The given examples, however, suffice for the purpose of this paper.

conditions that favour the emergence of compositionality. This assumption is adopted here too, and it is left for future research to investigate how such learning mechanisms could have evolved.

Early computational studies on the origins and evolution of language have focused on the formation of lexicons in systems with predefined meanings [19,30,33,43,65]. These were followed by experiments on lexicon formation in systems connected with the real world [10,38,47,49,50,55,57]. Meanwhile research on the emergence of syntactic structures, such as compositionality, also gained popularity. Usually these models have incorporated a predefined semantics [2,6,8,23]—or even no semantics at all [18,69]. The researchers of these 'ungrounded' studies have justified their design choice by saying that they only look at how syntactic structures can emerge and that this choice allows them to concentrate on that particular aspect. In a way they are right, because it will have helped them to focus on the emergence of syntax and their results are definitely insightful. However, it seems realistic to assume that the semantics of languages have arisen from a co-development of language and meaning in an embodied interaction of individuals with the real world, see, e.g., [16,27,66]. Omitting such interactions could have important consequences on the emergence of language in these individuals and hence in the population as a whole.

The world in which we have used and developed language is already highly structured. Apples, for instance, all have roughly the same shape, but some can have different colours. So, a red apple can be categorised based on its shape and colour as `object(shape(round),colour(red))`. In a way, one could say that parts of the object's regular properties map directly on parts of the semantic description of the object. By exploiting these regularities, e.g., based on statistical correlations, it appears plausible that such structures have been utilised in natural languages. This idea is elaborated upon in the current study in which it is assumed that syntactic structures co-evolve with semantic structures. More formally, the following twofold hypothesis is proposed:

(1) The emergence of compositional linguistic structures is based on exploiting regularities in (possibly random and holistic) expressions, cf. [67], though constrained by semantic structures.
(2) The emergence of combinatorial semantic structures is based on exploiting regularities found in the (interaction with the) natural world, though constrained by compositional linguistic structures.

This hypothesis is investigated using a simulated multiple robot model.

Up to now, only a few studies have used (simulated) robots as a research platform to investigate the emergence of compositional (and other grammatical) structures in language [10,31,32,46,58]. The current work is an extension of the work presented in [58], which is based on the Talking Heads experiment [49] and implements the language game model [43] in combination with the iterated learning model [6,23]. The model further incorporates machine learning techniques on grammar induction adapted from [14,54]. In the experiments the effect is investigated of evolving compositional languages following the iterated learning model, but with semantic development in connection to the simulated Talking Heads world. In particular, the effect on the language dynamic and its stability of using different

social strategies for language acquisition [55,62] and imposing a transmission bottleneck is studied.

The next section presents the simulation platform, the language game model and the iterated learning model. Section 3 then explains the grammar inducer used to evolve compositional structures. Experimental results are presented in Section 4 and these are discussed in Section 5. Finally, Section 6 concludes the paper.

## 2. Talking Heads

### *2.1. Modelling language evolution*

Before explaining the model, some things need to be said concerning the modelling of language evolution. Generally, the scientific aim of evolutionary linguistics is to study how modern languages have evolved from a stage prior to language. Traditionally, this field is studied by linguists, biologists, anthropologists, psychologists and primatologists. With the rise of the computer and advancements in Artificial Intelligence and Artificial Life, computational modelling became another methodology for studying language evolution. Models of language emergence and evolution typically involve a simulation using a multi-agent system of which the individuals are able to communicate, perceive their world and learn (often using standard machine learning techniques). They are typically situated in a world, which may be highly abstract (e.g., in studies where the meanings are predefined) or more realistic (e.g., in robotic models such as the one presented here). Modelling language evolution turns out to be very useful, because it "provides a complementary methodology that can help researchers to develop detailed and precise hypotheses on language origins and evolution and to test these hypotheses in the virtual experimental laboratory of the simulation" [11, p. 5].

The modelling studies can roughly be divided in three parts: (1) *origins*, (2) *emergence* and (3) *evolution*. The origins question investigates how and why humans (and other species) came to use communication, and how the language processing, creation and acquisition mechanisms have evolved. The emergence question assumes communication, processing, creation and acquisition mechanisms to be present and investigates how, given these mechanisms, languages or aspects thereof come about. Finally, given a language and the mechanisms mentioned, the evolution question investigates how the language then evolves (or changes) over time. The current study ignores the origins question and focuses on the emergence and evolution of language (or compositionality in particular).

Although the field investigates the evolution of *human* languages, using computer simulations means that the models are (often very crude) *simplifications* of human languages. Typically, the models focus on one aspect of language (e.g., the emergence of compositionality) and even then may abstract away from the natural human case. For instance, in the current model

- the language that emerges is not used for any other function than to describe objects; the individuals in the model do not use the communication to solve any other task than to learn from each other;

- interaction protocols are predefined and complex mechanisms, such as establishing joint attention or providing corrective feedback, are extremely simplified;
- no effort is being made to provide the agents with a realistic phonological system: words are invented as random strings of letters taken from a subset of the English alphabet and hence appear gibberish to humans;
- the agents evolve their language in an abstract environment (based on a real physical experiment) where they communicate about colours and shapes;
- the way agents perceive colours is far from realistic with respect to humans. Instead of perceiving colours in a realistic manner (e.g., through the widely accepted CIE *L*a*b** colour space), the agents perceive the colours on the computer based RGB-colour space. So, although colours appear prominently in this paper, the aim is not to investigate the evolution of colour terms in human language (see [47] for a related study on the emergence and evolution of human-like colour terms).

The reasons behind making such choices are that they allow straightforward implementation, while making a qualitative study on the specific research question, which in this case is: Given some means for communication, perception and learning, how can compositional structures in language emerge and evolve stably?

As mentioned, the current model is based on the Talking Heads experiment [3,49], which originally consisted of a set of agents that could embody themselves in a pan-tilt camera (thus becoming a physical robot) with which they could look at a scene displayed at a white board. The scene typically contained coloured objects of a certain shape about which the population tried to evolve a lexicon. The experiment was connected to the Internet, which allowed human users to interact with the system by, e.g., adding and removing agents to the population.

The current model is based on a simulation of this Talking Heads experiment and is part of the THSim toolkit [59]. This toolkit, including its source code, is freely downloadable from the Internet[4] and allows users to investigate many aspects of lexicon formation and to repeat the experiments described in this paper. THSim implements different versions of language games and the iterated learning model.

### *2.2. The language game*

In short, the model implements a situation in which the population of one generation transmits their language to the next generation by engaging in *language games* [43]. The language game (Fig. 1) is played by two agents: the speaker, which is typically an agent from the older generation, and the hearer, which is typically an agent from the new generation. In the language game, both agents perceive (`p`) a context $C$ containing a given number of objects. For each object, the agents extract features (`f`) that describe the objects' properties. The speaker selects one object as the *topic* (or target) and may inform the hearer non-verbally which object is the topic. This is similar to pointing. If the hearer is not informed, it considers all objects in the context as a potential topic.

[4] http://www.ling.ed.ac.uk/~paulv/thsim.html.

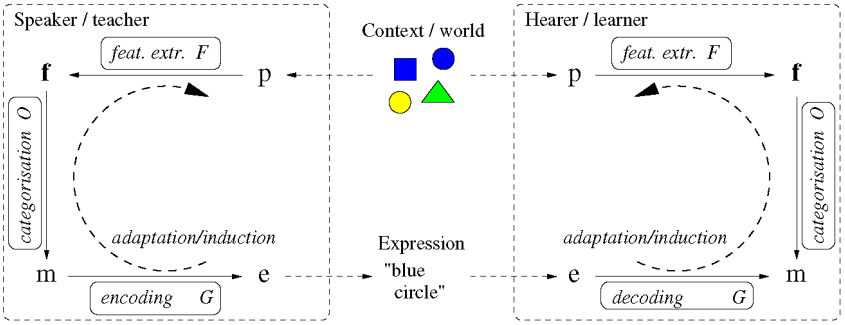


Fig. 1. The semiotic square illustrates the processes and used memory sources (ontology O and grammar G) of a language game between two agents: a speaker/teacher and a hearer/learner. See the text for details.

Both agents—individually—form a category (or meaning $\mathtt{m}$) that distinguishes the topic (or each potential topic) from the rest of the context. If no such category can be found, a new one may be constructed and added to the agents' private ontologies $\mathcal{O}$. The categorisation is modelled using the *discrimination game* model [44], explained in Section 2.5.

The speaker then tries to *encode* the topic's category into an *expression* ($\mathtt{e}$), which the hearer then tries to *decode*. If there is more than one way to encode or decode an expression, the agent selects the way that has the highest score, based on a number of weights that indicate the effectiveness of grammatical entries and their semantic properties. These weights are adapted according to the effectiveness of each game. If there is no way to encode or decode an expression, the agents construct new knowledge by inventing new rules or inducing rules from heard expressions. This new knowledge is then added to the agents' private grammars $\mathcal{G}$.

Two different types of language games are implemented for the current study: *observational games* and *guessing games*.[5]

*Observational game*: The speaker informs the hearer which object is the topic, thus establishing joint attention prior to the verbal communication. How this is done in real life is unimportant for the purpose of this paper, but children and caregivers seem to engage in joint attention quite frequently [52] and robots could do this using pointing [50]. Weights are adapted following Hebbian learning, as explained in Section 2.8.

*Guessing game*: The speaker does not inform the hearer about the topic, but the hearer has to guess which object is the topic, given the context and the utterance. The hearer then informs the speaker about its guess and the speaker provides the hearer with (corrective) feedback about whether the hearer guessed right or not. It is unclear

[5] In THSim, a third type of language game is implemented, which I have called the *selfish game* [55]. This game implements a *cross-situational statistical learner* [64], which is based on the cross-situational learner proposed by Siskind [37]. In cross-situational learning, the learner infers word-meanings from the co-variances of words with their meanings across different contexts/situations. Currently, this game has only been implemented to simulate lexicon formation.

whether children actually receive corrective feedback [4], but recent analysis indicate that they may [13]. Weights are adapted according to reinforcement learning.

These games differ mainly in the social strategy used to communicate, in particular with respect to the information about the topic handed over to the hearer. As a consequence of these different strategies, the learning mechanisms differ as well. For a comparison of these models with respect to lexicon formation, consult [55,62].

It is important to note that at the start of each agent $a$'s lifetime, its ontology $\mathcal{O}_a$ and grammar $\mathcal{G}_a$ are empty. Further, $\mathcal{O}_a$ and $\mathcal{G}_a$ are private representations and thus may differ from one agent to another. How the ontologies are represented, used and constructed is explained in Section 2.5. How the grammars are represented, used and induced is explained in Sections 2.6, 2.7 and 3. First, however, the iterated learning model is introduced and then the Talking Heads world is presented.

### *2.3. The iterated learning model*

The iterated learning model (ILM) [6,7,23,26] is a generic framework for simulating language evolution. The model (see Fig. 2) iterates over generations where the population is divided into two groups: *adults* and *learners*. The adults have passed the stage of learners and are assumed to have mastered the language. The learners enter the population as novice language users and acquire the language from scratch by observing the behaviour of adult speakers. The ILM cycles around iterations in which a given number of language games are played. At the end of each iteration, the adults 'die' and are replaced by the learners, who in turn are replaced by novel learners. This cycle then repeats. In short, the ILM involves a simplified model of a population dynamics and *vertical* transmission of language, i.e., the cultural transmission of language from one generation to the next where the output of one generation is the input to the next generation.

In most implementations of the ILM, including the current, the agents of all generations have the same 'phenotype', i.e., there is no biological adaptation but the languages evolve

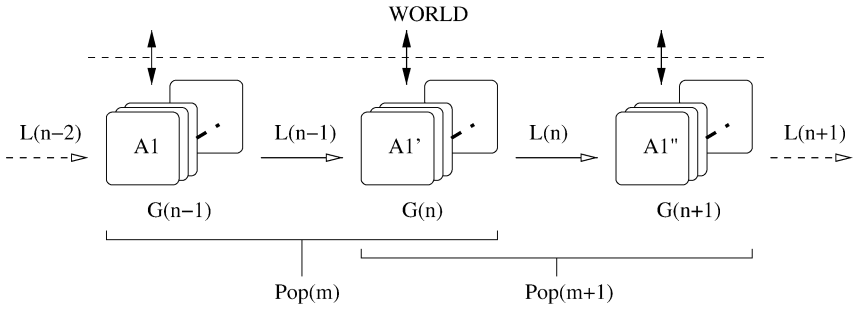


Fig. 2. A schematic view of the iterated learning model. A population of agents $A'$ from generation $G(n)$ acquires the language $L(n)$ by playing language games (as hearer) with agents $A$ from generation $G(n-1)$. After a given number of language games, generation $G(n-1)$ 'dies' and a new generation $G(n+1)$ is added to the population. Now generation $G(n)$ teaches this new generation and the cycle repeats.

culturally [6,23,69]. An exception in this respect is the work by Kenny Smith, who has studied the effect of biological adaptations on the learnability of lexicons [40].

### 2.4. The Talking Heads world

In the Talking Heads simulation (Fig. 3) a population of agents evolve language to communicate about their world. The world $\mathcal{W}$ consists of a set of geometrical coloured objects $o_i \in \mathcal{W}$, of which an arbitrarily selected subset of a fixed size appears at randomly selected locations in a display to form the *context* $C$ of a language game (upper left window in Fig. 3). In all experiments reported in this paper, the context size was fixed at 8 objects.

Each object is described with six different perceptible features: the (**r**)ed, (**g**)reen and (**b**)lue components of the **rgb** colour space, a (**s**)hape feature, and the (**x**) and (**y**)-coordinates of the objects' locations. In the current study, the agents only extract the first four features (**rgbs**). All features have real values between 0 and 1 and are designed such that they can be calculated by a real robot—in fact the features are highly similar to those used in the real Talking Heads experiment [49].

The shape feature $f_{\mathbf{s},i}$ of object $o_i$ is calculated as follows:

$$f_{\mathbf{s},i} = 2\frac{A_i}{A_{bb,i}} - 1 \tag{1}$$

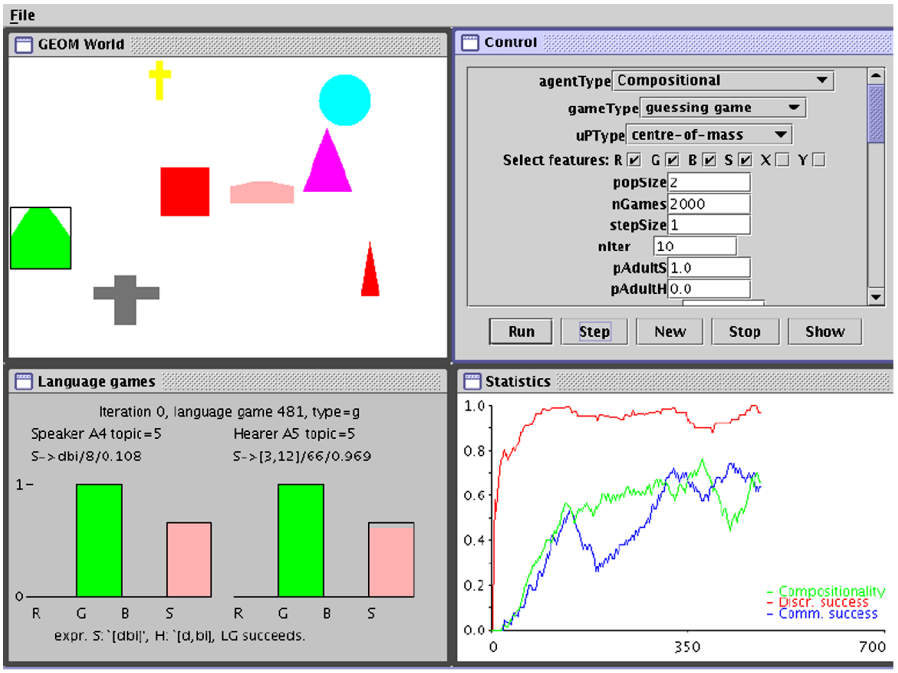


Fig. 3. A screen shot of the THSim toolkit. The upper left window shows the context of a language game, the upper right window contains the tool's control panel, the lower left window shows the details of a language game, and the lower right window shows some statistics of the simulation.

where $A_i$ is the area of object $o_i$ and $A_{bb,i}$ is the area of the smallest bounding box that can be drawn around object $i$. For example, a circle has a shape feature of $f_{\mathbf{s}} = 2\frac{\pi r^2}{(2r)^2} - 1 = \frac{\pi}{2} - 1$, squares have features $f_{\mathbf{s}} = 1$ and triangles features of $f_{\mathbf{s}} = 0$. The objects in the world are designed such that for two different shapes $\mathit{shape}_i \neq \mathit{shape}_j$ it holds that the shape features are different, i.e., $f_{\mathbf{s},i} \neq f_{\mathbf{s},j}$ (with the exception of squares and rectangles).

The world contains a total of 10 different shapes and 12 different colours, which are—more or less—basic colours with a non-uniform distribution in the **rgb** space, but well distinctive from each other. Each shape can be combined with each colour, hence the world has 120 different objects. The objective of the simulation is that the population develops a language from scratch to communicate successfully about their world.

### *2.5. The discrimination game*

Each agent $a$ constructs its private ontology $\mathcal{O}_a$ by playing discrimination games [44]. The objective of the discrimination game is to find one or more categories for an object (the topic) that distinguishes this topic from all other objects in the context. In case of failure, the agent expands its ontology in order to improve its discriminative ability for future games. The ontology contains the basic building blocks for constructing the categories (or meanings), which the agents use as an internal representation of the world's objects.[6] Categories are represented by prototypes $\mathbf{c} = (c_1, \ldots, c_n)$, which are points in an $n$-dimensional *conceptual space* [16]. The region in the space of which the points are nearest to a prototype is defined as the category of this prototype.

The dimensions of the conceptual space are called *quality dimensions* [16], and it is along these dimensions that the agents construct their ontology. The quality dimensions the agents use are directly related to the feature qualities the agents detect when seeing an object: **r**, **g**, **b** and **s**. With these qualities, the agents can construct various conceptual spaces, such as, for instance, a colour space using **rgb**, a shape space using dimension **s** or a 'redgreen' space using **rg**. For convenience, the spaces are denoted as a string indicating their dimensions, such as **rgbs** denotes the conceptual space of all available dimensions. For the sake of consistency, the meanings of sentences are assumed to cover all dimensions **rgbs**. The meanings are stored in a holistic conceptual space (spanned by all dimensions), or they can form a composition from other conceptual spaces, such as the colour space combined with the shape space. Meanings cannot be constructed from overlapping spaces, so combinations of **rgb** with **gs** are not allowed.

The basic units of the ontologies are called *categorical features* $c_i$. These are points in one dimension $i$, which segment dimension $i$ into areas of which the categorical features are the central points. Categories are constructed by combining the categorical features of different dimensions. Combining the dimensions leads to the formation of a conceptual space, as shown in Fig. 4.

[6] Sometimes the term *category* is used instead of *meaning*. For convenience, these are used to denote a representation. Obviously there are important differences between meanings and representations as I have discussed elsewhere [56].

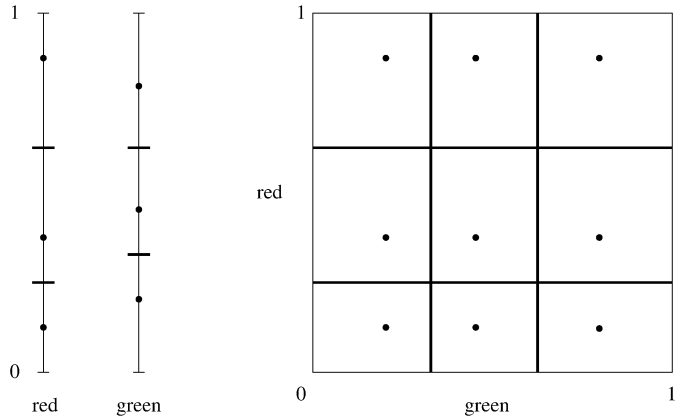


Fig. 4. This figure shows an example of how an ontology can be interpreted and used in terms of a conceptual space. The two quality dimensions **red** and **green**, each with three categorical features (points) form 3 regions in each dimension (left). When combined, these dimensions form a conceptual space that has 9 different categories (right). Note that the number of categorical features is not fixed, nor is the number of different conceptual spaces.

In a discrimination game, played by both the speaker and the hearer, each individual agent tries to distinguish the topic $o_t \in C$ from all other objects in the context $o_k \in C \backslash \{o_t\}$. (In a guessing game, the hearer plays a discrimination game for each object in the context.) Each agent $a$ categorises all objects in the context $o_j \in C$ by taking for each feature $f_{i,j}$ the categorical feature $c_i \in \mathcal{O}_a$ that is nearest to $f_{i,j}$, and then combining each dimension to form the category $\mathbf{c}_j = (c_1, \ldots, c_n)$. The topic's category $\mathbf{c}_t$ is *distinctive* if it is not a category for any other object in the context, i.e., $\forall o_j \in C\colon o_j \neq o_t \Rightarrow \mathbf{c}_j \neq \mathbf{c}_t$.

If the topic's category is a *distinctive category*, it becomes the topic's *meaning* $m_t$. Otherwise, the discrimination game fails—and consequently the whole language game likewise—and the agent adds each of the topic's features as exemplars of new categorical features to its ontology $\mathcal{O}_a$, unless the exemplar already exists in $\mathcal{O}_a$.

One further remark on notation is required at this stage. Assuming the 4 quality dimensions used in the current implementation, I denote an object's category as a 4-dimensional vector with the order **rgbs**. For example, a red square is written as $(1, 0, 0, 1)$. The meanings from lower dimensional conceptual spaces are denoted differently. For example the colour red is denoted by the list $[1_{\mathbf{r}}, 0_{\mathbf{g}}, 0_{\mathbf{b}}]$ and the shape square is denoted by $[1_{\mathbf{s}}]$. I use this distinction for convenience. The vector notation is a more common notation and stands for the meaning as a whole. The list notation allows us to form more complex meanings using its elements and combine them with possibly other lists, as will be described shortly.

### *2.6. The grammar*

Each agent $a$ constructs its private grammar $\mathcal{G}_a$ ontogenetically, starting from an initially empty set. The grammar is defined as a set of rules $R$ that are schemas (or rewrite rules), which may be either holistic or compositional. Table 1 provides an example grammar. In

Table 1
An imaginary grammar used in the examples of the text. The rules are realistic with respect to the model, apart from the strings which are based on English. In the model, strings are constructed from random letters taken from a subset of the English alphabet. See the text for a detailed explanation

| | | |
|---|---|---|
| $R_1$: | S/**rgbs** $\rightarrow$ A/**rgb** B/**s** | 0.85 |
| $R_2$: | S/**rgbs** $\rightarrow$ B/**s** A/**rgb** | 0.06 |
| $R_3$: | S/**rgbs** $\rightarrow$ C/**rb** D/**gs** | 0.1 |
| $R_4$: | S/**rgbs** $\rightarrow$ redsquare/$[1^{0.25}_{\mathbf{r}}, 0^{0.2}_{\mathbf{r}}, 0^{0.25}_{\mathbf{g}}, 0^{0.2}_{\mathbf{b}}, 1^{0.1}_{\mathbf{s}}]$ | 0.25 |
| $R_5$: | A/**rgb** $\rightarrow$ blue/$[0^{0.2}_{\mathbf{r}}, 0^{0.2}_{\mathbf{g}}, 1^{0.2}_{\mathbf{b}}]$ | 0.2 |
| $R_6$: | A/**rgb** $\rightarrow$ yellow/$[1^{0.9}_{\mathbf{r}}, 1^{0.9}_{\mathbf{g}}, 0^{0.9}_{\mathbf{b}}]$ | 0.9 |
| $R_7$: | B/**s** $\rightarrow$ triangle/$[0^{0.8}_{\mathbf{s}}]$ | 0.8 |
| $R_8$: | B/**s** $\rightarrow$ double/$[0^{0.01}_{\mathbf{s}}]$ | 0.01 |
| $R_9$: | C/**rb** $\rightarrow$ rue/$[1^{0.1}_{\mathbf{r}}, 1^{0.1}_{\mathbf{b}}]$ | 0.1 |
| $R_{10}$: | D/**gs** $\rightarrow$ greenagon/$[1^{0.1}_{\mathbf{g}}, 0.5^{0.1}_{\mathbf{s}}]$ | 0.1 |
| $R_{11}$: | D/**gs** $\rightarrow$ double/$[0.5^{0.01}_{\mathbf{g}}, 0^{0.01}_{\mathbf{s}}]$ | 0.01 |

this grammar, S, A, B, C and D are syntactic categories.[7] The rules indicate how the left hand sides rewrite to the right hand sides. Some rules ($R_1$–$R_3$) rewrite to compositional rules and the others rewrite to single word utterances. The bold-face strings after the slashes indicate the conceptual space which is *covered* by the syntactic category.

By definition, all sentences S cover all dimensions of the entire conceptual space **rgbs**; the other syntactic categories cover a space of lower dimensionality, such as colour **rgb** (A), shape **s** (B), 'redblue' **rb** (C) and 'greenshape' **gs** (D). The terminal slots (rules $R_4$–$R_{11}$) rewrite to single words, such as "redsquare" and "blue". (Note that $R_4$ is a holistic rule.) Sentences are defined to cover all dimensions in order to reduce the search space of possibilities.

Compositions can be formed by applying terminal slots of the appropriate syntactic category to compositional rules (using the composition operator $\circ$). For instance, rules $R_5$ and $R_7$ can be applied to rule $R_1$ to form the composition $R_1 \circ R_5 \circ R_7$, which rewrites to the sentence "bluetriangle". In addition, we can apply the same terminal slots to rule $R_2$ to form the composition $R_2 \circ R_7 \circ R_5$ obtaining "triangleblue". Both compositions have the meaning $(0, 0, 1, 0)$. The meaning of a rule (and consequently the composition) is formed from the categorical features listed between square brackets, as described below.

The categorical features are denoted as a value with its dimension in subscript and its *category weight* $w(c_i)$ in superscript. For instance, $1^{0.25}_{\mathbf{r}}$ is a categorical feature in the quality dimension for red (**r**) with a weight of $w(1_{\mathbf{r}}) = 0.25$. The features can be combined to construct a meaning. Each possible meaning of $R_4$ must cover the holistic conceptual space of **rgbs**, so they must include a categorical feature from each dimension.

[7] Although the rules make the grammar look context-free, it is not because in the current implementation, all non-terminals of the sentence (i.e., A, B, C and D in the example) rewrite to terminals. Hence, the grammar is finite. Strictly speaking, the grammar is also not regular, since the top rules can contain two non-terminals, but the grammar could be represented by a regular grammar. Future research aims to extend the current model to grammars that are context-free.

The two possible meanings that can be formed from the list $[1^{0.25}_{\mathbf{r}}, 0^{0.2}_{\mathbf{r}}, 0^{0.25}_{\mathbf{g}}, 0^{0.2}_{\mathbf{b}}, 1^{0.1}_{\mathbf{s}}]$ are thus $m'_4 = (1, 0, 0, 1)$ representing a red square and $m''_4 = (0, 0, 0, 1)$ representing a black square. We can calculate the *meaning weights* $w$ for these meanings as the average of the category weights used in the meaning. So, meaning $m'_4$ has meaning weight $w'_4 = 0.20$ and $m''_4$ has meaning weight $w''_4 = 0.19$. In general, the meaning weights are calculated as the average weight of the categorical features used in the meaning, i.e.,

$$w_i = \frac{1}{d_i} \sum_{j=1}^{d_i} w(c_{j,k}) \tag{2}$$

where $d_i$ is the dimension of the rule's covering conceptual space. The category weights $w(c_{j,k})$ are updated to indicate how well they have been used to construct the rule's meaning (Section 2.8). So, in the example, rule $m'_4$ represents $R_4$ better than $m''_4$, because $0.20 > 0.19$. The agents use these weights to learn the proper meaning of a rule, based on a selectionist competition.

Each rule $R_i$ further has a *rule weight* $\rho_i$, as given in the final column of Table 1, which indicates the effectiveness of the rule in previous language games. Given these weights one can calculate the score $s_i$ of rule $R_i$ by:

$$s_i = \begin{cases} \rho_i w_i & \text{if } R_i \text{ has specified meanings,} \\ \rho_i & \text{if } R_i \text{ rewrites into non-terminals.} \end{cases} \tag{3}$$

If we have a composition $\bigoplus R_i$ to form a sentence with one or more components $R_i$, we can attribute a score to this composition. This score $\sigma(\bigoplus R_i)$ is the product of the rules' scores $s_i$, i.e.,

$$\sigma\left(\bigoplus R_i\right) = \prod_{\bigoplus R_i} s_i \tag{4}$$

where $\bigoplus R_i$ is a shorthand for the composition. Suppose we have the composition $R_1 \circ R_5 \circ R_7$ to mean (0,0,1,0), then

$$\begin{aligned} \sigma(R_1 \circ R_5 \circ R_7) &= s_1 \cdot s_5 \cdot s_7 = (0.85) \cdot \left(0.2 \cdot \frac{0.2 + 0.2 + 0.2}{3}\right) \cdot \left(0.8 \cdot \frac{0.8}{1}\right) \\ &= 0.02176. \end{aligned}$$

In contrast, the competing composition $R_2 \circ R_7 \circ R_5$, with the same meaning, has score

$$\sigma(R_2 \circ R_7 \circ R_5) = (0.06) \cdot \left(0.2 \cdot \frac{0.2 + 0.2 + 0.2}{3}\right) \cdot \left(0.8 \cdot \frac{0.8}{1}\right) = 0.001536.$$

The score $\sigma(\bigoplus R_i)$ is used to select among competing rules when encoding or decoding an expression.

The arithmetic mean is used for calculating the meaning weights and the product is used when calculating the scores, in an analogy with probabilities. In this analogy, the selection of category features is considered as an independent process (hence summation), but the construction of compositions is not (hence multiplication).

### 2.7. Encoding and decoding

Encoding and decoding are straightforward techniques to match the meaning of the language game's topic with a set of possible compositions or to match an expression and (possible) meaning with a set of compositions. When there are multiple ways of encoding or decoding an expression, then the composition with the highest score $\sigma(\bigoplus R_i)$ is selected. Encoding and decoding can be described as follows:

*Encoding*: If a speaker/teacher tries to produce an expression, it searches for all possible compositions in its grammar that match the meaning. For example, if the speaker tries to encode an utterance for meaning (0,0,1,0) using the grammar shown in Table 1, it will come up with compositions $R_1 \circ R_5 \circ R_7$ and $R_2 \circ R_7 \circ R_5$. Based on the scores calculated as described in Eq. (4) above, the speaker will select the highest scoring composition and produces the utterance "bluetriangle".

*Decoding*: If a hearer receives an utterance, it will interpret the expression. (Utterances are received without word boundaries, so there may be different ways to segment a sentence.) The hearer creates a temporary list by searching its grammar for compositions that decode the utterance and removes from this list all compositions whose semantics do not match any of the possible meanings in the given language game. In the current implementation of the observational game, there is only one possible meaning, but in the guessing game there are typically more.

Suppose, for example, the hearer receives the utterance "bluetriangle" in an observational game with the topic's meaning $(0, 0, 1, 0)$. Further assume that the hearer has the grammar of Table 1. The utterance is then interpreted with composition $R_1 \circ R_5 \circ R_7$, which the hearer selects as interpretation. If there are more possible parses, the hearer selects the one with the highest score $\sigma(\bigoplus R_i)$. In this example, the observational game is a success, because the hearer found an interpretation matching the topic's meaning. If the topic has a different meaning, the observational game fails.

Now suppose that the hearer receives the utterance "redsquare" in a guessing game with a context of three categorised objects: a red square with meaning $m_1 = (1, 0, 0, 1)$, a black square with $m_2 = (0, 0, 0, 1)$ and a yellow square $m_3 = (1, 1, 0, 1)$. The utterance can be interpreted in two ways, both using the holistic composition $R_4$. The interpretation matching $m_1$ has a score of $\sigma(R_4)' = 0.050$ and the one matching $m_2$ yields $\sigma(R_4)'' = 0.047$. The hearer selects $m_1$ as the topic's meaning and if this topic (the red square) is the object intended by the speaker, the guessing game succeeds, otherwise there is a mismatch in reference and the game fails.

### 2.8. Adaptation of weights

When the language game has finished, the effectiveness of the game is evaluated. Depending on the outcome, both agents adapt the weights which they use to calculate the scores:

*Success*: If the language game is a success, both the speaker and hearer increase the weights $\rho_i$ of the rules $R_i$ that are part of the composition according to:

$$\rho_i = \eta \cdot \rho_i + 1 - \eta \tag{5}$$

where $\eta = 0.9$ is a constant learning parameter. In addition, rules $R_j$ that are part of a competing composition are laterally inhibited following:

$$\rho_j = \eta \cdot \rho_j. \tag{6}$$

Furthermore, the weights of the categorical features that constitute the meaning of the successfully used rules are increased by

$$w(c_{i,k}) = \eta \cdot w(c_{i,k}) + 1 - \eta. \tag{7}$$

The competing categorical features in the same rules and the categorical features that constitute competing rules are inhibited:

$$w(c_{j,k}) = \eta \cdot w(c_{j,k}). \tag{8}$$

*Mismatch*: (Guessing game only.) In case of a mismatch, only the hearer decreases the weights of the rules and its used categorical features according to:

$$\rho_i = \eta \cdot \rho_i \tag{9}$$

and

$$w(c_{i,k}) = \eta \cdot w(c_{i,k}). \tag{10}$$

Note that the lateral inhibition of the rule weight $\rho_j$ helps to disambiguate different grammatical structures, while the lateral inhibition of the categorical features' weights $w(c_j, k)$ helps to disambiguate the meaning of a rule. Rules are said to *compete* if they are part of a composition that parsed the distinctive category (speaker) or utterance (hearer), but were not selected.

The equations ensure that the weights remain between 0 and 1, while the choice of $\eta = 0.9$ allows the weights to maintain a rather long history of past experiences. The choice of this update rule is not extremely important (Steels and Kaplan, for instance, use a different update rule [48]), but it has been found that the current equations work better than when associations are updated using a frequency counter, together with a Bayesian learning model. This is mainly because with the currently used functions, the scores fluctuate faster, thus strengthening competition, than would be the case in a probabilistic model.

## 3. Grammar induction

If the speaker cannot produce an utterance or if the hearer fails to interpret an utterance, the agent in question has to expand its linguistic knowledge. The speaker may *invent* new knowledge and the hearer may induce new knowledge. New compositional structures can only be constructed by the hearer.

### 3.1. Speaker's invention

If a speaker is not able to produce an utterance, the grammar is insufficient to encode the meaning and new knowledge has to be invented. This is done in one of the two following ways: *exploitation* and *holistic creation*. These invention mechanisms are similar to those used in [23,24].

*Exploitation*: The speaker does not invent new compositional rules, but can exploit an existing rule if this rule is able to encode a part of the sentence. For instance, if a speaker with the grammar of Table 1 wishes to produce an utterance to express the meaning $(0, 0, 1, 1)$, then both the compositions $R_1 \circ R_5 \circ ?$ and $R_2 \circ ? \circ R_5$ produce a partial encoding covering the meaning part $[0_{\mathbf{r}}, 0_{\mathbf{g}}, 1_{\mathbf{b}}]$. The speaker then invents a new rule in which it associates the complement of the meaning, $[1_{\mathbf{s}}]$ with a newly constructed word, such as "rectangle". This leads to the construction of the new rule:

$$R_{12}: \quad \mathbf{B/s} \rightarrow \mathbf{rectangle}/[1_{\mathbf{s}}^{0.01}] \quad 0.01.$$

This rule is then immediately applied to utter the expression "bluerectangle" using composition $R_1 \circ R_5 \circ R_{12}$, which has the highest score ($\sigma(R_1 \circ R_5 \circ R_{12}) = 3.4 \times 10^{-6}$ and $\sigma(R_2 \circ R_{12} \circ R_5) = 2.4 \times 10^{-7}$).

*Holistic creation:* If the speaker cannot encode *any* part of the topic's meaning, then a holistic rule is created. For instance, no part of meaning $(0, 1, 1, 1)$ can be encoded with the grammar given in Table 1. Hence, a new word is created and the association is added to the grammar. For example, the speaker might add rule

$$R_{12}: \quad \mathbf{S/rgbs} \rightarrow \mathbf{cyanrectangle}/[0_{\mathbf{r}}^{0.01}, 1_{\mathbf{g}}^{0.01}, 1_{\mathbf{b}}^{0.01}, 1_{\mathbf{s}}^{0.01}] \quad 0.01$$

to its grammar. This rule is immediately incorporated to produce the utterance "cyanrectangle".

In the above examples, words are constructed based on English for illustrative purposes only. In the simulation, words are constructed as random strings of letters from a given alphabet $\Sigma$. The strings have a length $2 \leqslant l \leqslant 8$ where the length follows a probability distribution of $f(l) \propto 1/l$, which is typical for human languages [68]. When new rules or meanings are added to the grammar, all weights are initialised with value $w(c) = \rho = 0.01$.

### 3.2. Interpretation and induction

When the hearer is not able to decode an utterance, the hearer has to induce the utterance's meaning, such that it is consistent with the speaker's intention. In case of the observational game the hearer already knows the topic, in the case of the guessing game the speaker will now hand over this information. The core of our interest now lies in the grammar induction, which at this stage is the same for both the observational and guessing game.

In modern psycholinguistics it is widely assumed that humans learn grammar from observing other humans' linguistic behaviours, e.g., [28,29,53], rather than by tuning parameters in relation to some innate universal grammar as proposed by Chomsky [12].

According to Tomasello, Lieven and co-workers, analysis of interactions between mothers and children reveal that children seem to gradually construct grammatical schemas based on similarities (or *alignments*) found in different sentences used by their caregivers.

Consider, for example, the following two sentences:

(1) I love Mum.
(2) I love Maria.

If a learner hears these two phrases for the first time, then the learner can infer that both 'Mum' and 'Maria' belong to the same linguistic category, and that the segment 'I love' belongs to something else.

This type of learning has been implemented computationally as an *alignment-based learner* (ABL) [54], where the learner induces grammatical structures from sentences stored in linguistic corpora. This learner is based on alignment learning combined with some selection criterion. Van Zaanen has shown that calculating the probability of hypotheses based on past experiences yields a good selection criterion. The current model implements ABL, but incorporates a selection criterion based on the weights as explained in the previous section. (Kirby's models [22–24], too, uses alignment learning, but his more recent models [23,24] are without any selection criteria.)

Basically, there are three different induction mechanisms: *exploitation*, *chunking* and *incorporation*. These induction mechanisms are followed by an additional step called *generalise and merge* explained in Section 3.3.

*Exploitation*: Exploitation is used when the learner is capable of decoding only a part of the sentence. In that case, the learner adds a new rule to cover the remaining part of the sentence, similar to the speaker's exploitation rule. Again consider the grammar of Table 1. When the hearer receives the expression "bluerectangle" to mean $(0, 0, 1, 1)$, this parses partially to the composition $R_1 \circ R_5 \circ ?$, covering meaning $[0_{\mathbf{r}}, 0_{\mathbf{g}}, 1_{\mathbf{b}}]$. In this case a new rule is constructed with non-terminal B as its head, the word "rectangle" and meaning $[1_{\mathbf{s}}^{0.01}]$; i.e., rule

$$R_{12}: \quad \mathbf{B/s} \rightarrow \mathbf{rectangle}/[1_{\mathbf{s}}^{0.01}] \quad 0.01$$

is added to the learner's grammar. (Note that the utterance does not partially decode into $R_2 \circ ? \circ R_5$ because of the wrong word order in this composition.)

*Chunking*: Chunking is incorporated when an utterance-meaning pair is not (partially) parseable, but when a part of the utterance-meaning pair aligns with stored holistic rules, or more precisely with stored utterance-meaning pairs. In each game where the hearer receives an utterance and successfully categorises the topic, it stores the utterance-meaning pair in the set of instances $\mathcal{I}$, such as shown in Table 2. When an utterance cannot be decoded or exploited, then the learner searches alignments between utterances and stored instances. For each found alignment, the learner also checks for alignments at the semantic level. For the instances that have alignments at the semantic and utterance level, the learner keeps track of the frequency with which these instances have been observed previously. The learner then decides to make a split where it appears most effective based on the frequencies or,

Table 2
An example list of stored instances of utterances (1st column), their meanings (2nd column) and their frequencies (3rd column)

| | | |
|---|---|---|
| redsquare | (1, 0, 0, 1) | 3 |
| bluetriangle | (0, 0, 1, 0) | 3 |
| yellowtriangle | (1, 1, 0, 0) | 3 |
| triangleyellow | (1, 1, 0, 0) | 1 |
| greendouble | (0, 1, 0, 0) | 2 |
| ruegreenagon | (1, 1, 1, 0.5) | 2 |
| ruedouble | (1, 0.5, 1, 0) | 1 |

in case of a tie, based on the largest common chunk. It then will add the new rules to the grammar as illustrated in the following example. (Note that the incorporation of an instance-base deviates from Van Zaanen's ABL and is inspired by the memory-based learning techniques, which uses nearest neighbourhood classification, as used in [14].)

Suppose, for example, the hearer receives the word "redcircle" with meaning (1, 0, 0, 0.57). The following alignments will then be found in Table 2.

| | $f$ | instance | | utterance | |
|---|---|---|---|---|---|
| a. | 3 | redsquare | (1, 0, 0, 1) | redcircle | (1, 0, 0, 0.57) |
| b. | 3 | redsquare | (1, 0, 0, 1) | redcircle | (1, 0, 0, 0.57) |
| c. | 3 | bluetriangle | (0, 0, 1, 0) | redcircle | (1, 0, 0, 0.57) |
| d. | 3 | yellowtriangle | (1, 1, 0, 0) | redcircle | (1, 0, 0, 0.57) |
| e. | 2 | ruegreenagon | (1, 1, 1, 0.5) | redcircle | (1, 0, 0, 0.57) |
| f. | 1 | ruedouble | (1, 0.5, 1, 0) | redcircle | (1, 0, 0, 0.57) |
| g. | 1 | ruedouble | (1, 0.5, 1, 0) | redcircle | (1, 0, 0, 0.57) |

Two notes are necessary here: 1) Alignments can only be made either at the start or at the end of an utterance, so alignments "redcircle", "redcircle" and "redcircle" are not allowed. This restriction does not hold at the semantic level. For example, the semantic alignment between (1, 0, 1, 0) and (1, 1, 1, 1) is valid. With this restriction, it is assumed that connected parts in the string belong to one part of the meaning, which in future models may be further decomposed, and—above all—it prevents the emergence of meaningless substrings as was the case in Kirby's models [23,24]. And 2) "greendouble" is also not taken, because there is no alignment in the semantics.

The selection of which chunks are made now follows two criteria:

(1) The frequencies of identical chunks in the utterance are summed. This yields three possible chunks with a maximum total frequency of 3:
   (i) redcircle with (1, 0, 0, 0.57) according to (a),
   (ii) redcircle with (1, 0, 0, 0.57) according to (b), and
   (iii) redcircle with (1, 0, 0, 0.57) according to (e) and (g).
(2) The chunk that has the largest syntactic alignment is selected.

So, given these restrictions the hearer will choose to chunk the pair "redcircle" $(1, 0, 0, 0.57)$ according to line (a.) and the learner adds the following rules to its grammar:

$$R_{12}: \quad \mathbf{S/rgbs} \rightarrow \mathbf{E/rgb} \quad \mathbf{F/s} \quad 0.01$$
$$R_{13}: \quad \mathbf{E/rgb} \rightarrow \mathbf{red}/[1_{\mathbf{r}}^{0.01}, 0_{\mathbf{g}}^{0.01}, 0_{\mathbf{b}}^{0.01}] \quad 0.01$$
$$R_{14}: \quad \mathbf{F/s} \rightarrow \mathbf{square}/[1_{\mathbf{s}}^{0.01}] \quad 0.01$$
$$R_{15}: \quad \mathbf{F/s} \rightarrow \mathbf{circle}/[0.57_{\mathbf{s}}^{0.01}] \quad 0.01.$$

This will lead to some redundancy, because $R_{12}$ is – apart from the non-terminal labels—the same rule as $R_1$. However, this will be repaired in the generalise and merge step described shortly.

*Incorporation*: The incorporation is done when no compositional structure can be induced. In this case, the utterance-meaning pair is adopted holistically. For instance the reception of utterance "lightgraycircle" with meaning $(0.25, 0.25, 0.25, 0.57)$ would result in the incorporation of:

$$R_{12}: \quad \mathbf{S/rgbs} \rightarrow \mathbf{lightgraycircle}/[0.25_{\mathbf{r}}^{0.01}, 0.25_{\mathbf{g}}^{0.01}, 0.25_{\mathbf{b}}^{0.01}, 0.57_{\mathbf{s}}^{0.01}] \quad 0.01$$

It is important to note that in the instance-base $\mathcal{I}$ all instances are stored the way they are received. This is in contrast to the grammar, which stores the way they can be rewritten; thus the instance-base allows the learner to search all instances they way they have been heard, which aids in finding an effective way to chunk up utterances based on past experiences, while not being restricted by previously made compositions which may not be effective. Moreover, the instance-based alignment learning can be used to 'unlearn' a previously learnt rule, when—based on later instances—other rules can describe the data better. Although it is unclear as to what extent humans store both whole utterances as exemplars and generalisations of these, there is growing evidence that both types of storage are used [29]. Most convincing evidence of this dual storage is found in relation with the storage of morphologically complex words [1].

These induction mechanisms differ from Kirby's model [23,24] mainly in four ways. First, the exploitation rule is an implicit property of Kirby's chunking mechanism. Second, Kirby's chunking mechanism allows alignments to be unconnected (as in reddouble), which, as mentioned, can lead to the emergence of substrings that have no semantic content, but it can also lead to an explosion of string length [42]. Third, the current chunking mechanism uses an instance-base for finding effective chunks. In Kirby's model, learners are restricted by previously made constructions, so they cannot unlearn a previously learnt compositions. This is also partly caused by the fourth difference with Kirby's model, where new rules subsume old ones which are deleted from the grammar. In this model, all rules are memorised, even the old ones, which allows the adaptation of weights to serve as a competitive selection mechanism, giving way for a *self-organisation* of language [45].

### 3.3. Generalise and merge

When new rules are induced, there may emerge some redundancies and other side effects that need to be dealt with. In order to deal with this, two post-operations have been introduced: *generalise* and *merge*.

*Generalise*: The generalisation step serves to exploit more regularities than was done in the chunking step. For example, suppose an agent has the following two rules in its grammar:

$$R_1\text{:}\quad \mathbf{S/rgbs} \rightarrow \mathbf{bluecircle}/[0^{0.2}_{\mathbf{r}}, 0^{0.2}_{\mathbf{g}}, 1^{0.2}_{\mathbf{b}}, 0.57^{0.2}_{\mathbf{s}}] \quad 0.2$$

$$R_2\text{:}\quad \mathbf{S/rgbs} \rightarrow \mathbf{yellowsquare}/[1^{0.5}_{\mathbf{r}}, 1^{0.5}_{\mathbf{g}}, 0^{0.5}_{\mathbf{b}}, 1^{0.5}_{\mathbf{s}}] \quad 0.5.$$

Further suppose that the utterance-meaning pair "bluesquare"-$(0, 0, 1, 1)$ was chunked with $R_2$—e.g., because its aligning instances had a higher occurrence frequency—so that the following rules were added:

$$R_3\text{:}\quad \mathbf{S/rgbs} \rightarrow \mathbf{A/rgb}\quad \mathbf{B/s} \quad 0.01$$

$$R_4\text{:}\quad \mathbf{A/rgb} \rightarrow \mathbf{blue}/[0^{0.01}_{\mathbf{r}}, 0^{0.01}_{\mathbf{g}}, 1^{0.01}_{\mathbf{b}}] \quad 0.01$$

$$R_5\text{:}\quad \mathbf{A/rgb} \rightarrow \mathbf{yellow}/[1^{0.01}_{\mathbf{r}}, 1^{0.01}_{\mathbf{g}}, 0^{0.01}_{\mathbf{b}}] \quad 0.01$$

$$R_6\text{:}\quad \mathbf{B/s} \rightarrow \mathbf{square}/[1^{0.01}_{\mathbf{s}}] \quad 0.01.$$

The agent missed the opportunity to chunk rule $R_1$. This is fixed by the generalisation step, which will add the rule

$$R_7\text{:}\quad \mathbf{B/s} \rightarrow \mathbf{circle}/[0.57^{0.01}_{\mathbf{s}}] \quad 0.01$$

to the grammar as well. Effectively, the generalisation step chunks all possible rules that fits the initial chunk.

*Merge*: The merge step is adapted from Kirby's model [23,24] and serves to reduce redundancy in the grammar. Two types of merging are applied. First, rules with different non-terminal labels that are effectively the same are merged. For instance, in the 'chunking' example of the previous subsection, the following rules were induced.

$$R_{12}\text{:}\quad \mathbf{S/rgbs} \rightarrow \mathbf{E/rgb}\quad \mathbf{F/s} \quad 0.01$$

$$R_{13}\text{:}\quad \mathbf{E/rgb} \rightarrow \mathbf{red}/[1^{0.01}_{\mathbf{r}}, 0^{0.01}_{\mathbf{g}}, 0^{0.01}_{\mathbf{b}}] \quad 0.01$$

$$R_{14}\text{:}\quad \mathbf{F/s} \rightarrow \mathbf{square}/[1^{0.01}_{\mathbf{s}}] \quad 0.01$$

$$R_{15}\text{:}\quad \mathbf{F/s} \rightarrow \mathbf{circle}/[0.57^{0.01}_{\mathbf{s}}] \quad 0.01.$$

$R_{12}$ is a copy of rule $R_1$ in Table 1, only with different labels on the non-terminals. The merging step merges these two rules (i.e., $R_{12}$ is removed) and then renames the labels E and F into A and B resp. for rules $R_{13}$–$R_{15}$. Second, rules that have the same non-terminal labels (or even different ones) and the same word-forms, but with different meanings covering the same conceptual space are merged. For example, the rules

$$R_i\text{:}\quad \mathbf{E/s} \rightarrow \mathbf{square}/[1_{\mathbf{s}}^{0.21}] \quad 0.21$$

$$R_j\text{:}\quad \mathbf{F/s} \rightarrow \mathbf{square}/[0.8_{\mathbf{s}}^{0.01}] \quad 0.01$$

are merged into:

$$R_i\text{:}\quad \mathbf{E/s} \rightarrow \mathbf{square}/[1_{\mathbf{s}}^{0.21}, 0.8_{\mathbf{s}}^{0.01}] \quad 0.22$$

where the scores of the two rules are added.

## 4. Experimental results

With the above models, three conditions were investigated whose results are presented in this section without any additional discussions and clarifications; these are presented in the next section. The first experiment is used as a baseline experiment. In this experiment, the observational game and the guessing game were run with a population size of 2 (1 adult and 1 learner). Although this is not a realistic population size, it allows us to investigate the basic behaviour of the model and is a typical setting for most ILM studies, e.g., [6,23]. Each agent, including the agents of the first iteration, started with an empty ontology and grammar. In the current experiments, the adult of the first iteration constructed the first version of the language, which then evolved over the iterations. As the speakers cannot invent compositional structures, the language of the first adult was holistic.

The second set of experiments investigates the effect of increasing the population size from 2 to 6. In this experiment, the population contained 3 adults and 3 learners, where the speakers of the language games are selected from the adult population and the hearers from the learner population. This set of experiments was carried without imposing a transmission bottleneck.

In the third set of experiments, a bottleneck on the transmission of language was imposed, both with a population size of 2 and 6. A bottleneck on the transmission means that each learner learns from its teachers by observing only a part of the teacher's language. When the learners becomes adults, they teach the next generation over *another* part of the language. In this set of experiments, a bottleneck was imposed by selecting a subset of the world as the set of objects from which contexts were selected. This subset differed in each learning episode (or iteration), but its size remained fixed at 50% (i.e., the agents in each iteration only observed 60 of the 120 objects). It has been shown in various studies with the ILM that imposing a bottleneck on the transmission of the language provides a strong external pressure on the emergence of compositional structures [6,23,41,69].

For all conditions, simulations were done with the observational game (OG) and the guessing game (GG). The simulations were run for 250 iterations of $N$ language games each, where $N = 2{,}500$ if the population size is 2, and $N = 6{,}000$ if the population size is 6. The $N$ language games within an iteration form the *training phase*. At the end of each iteration, before the adults were replaced by the learners and new learners entered the population, all agents were tested on a number of situations for certain aspects of their communicative ability. In this *test phase*, 200 language games were played with the adaptation and induction turned off, and where in each game each agent encodes an utterance

to convey the reference of one given topic and where each agent decodes the utterances expressed by the other agents. Whether or not a transmission bottleneck was imposed during the training phase, the test phase was always done with all 120 objects.

For each experiment, 10 different trials were run with different random seeds. Some results are presented as averages over the 10 different trials and—where necessary—together with their standard deviations. The results will be presented using 4 different measures calculated during the testing phase: *compositionality*, *production coherence*, *interpretation accuracy* and *similarity*.

*Compositionality*: The proportion of expressions that were encoded or decoded using compositional rules during the testing phase.

*(Production) coherence*: The fraction of agents that produced the same utterance to name objects during the testing phase, averaged over the 200 games played during this phase. (Note that this measure disregards whether the agents used the same grammatical rules.)

*(Communicative) accuracy*: The fraction of agents that could successfully interpret the produced utterances of the other agents in the population, averaged over the number of games played during the testing phase.

*Similarity*: The average proportion of the grammars of adults that are acquired by the learners at the end of an iteration. This is based on internal inspection, rather than on the testing situations.

All measures, except similarity, are reported graphically.

### *4.1. Baseline experiments*

The first set of experiments investigated the behaviour of both the OG and GG with a population size of 2 and without imposing a transmission bottleneck.

Fig. 5 shows the results of the baseline experiments for the OG (top) and the GG (bottom). Compositionality (graphs on the left) rose in both experiments swiftly to levels around 0.8. For the OG, this occurred already after two iterations; for the GG this took about 10 iterations. After that, compositionality in the OG fluctuated in general between 0.7 and 0.9, but with quite some drastic catastrophes where compositionality nearly disappeared for short periods. Compositionality evolved to an average of $0.73 \pm 0.24$ at the end of the final iteration. This means that at some points the languages changed from compositional languages into holistic ones, though each time compositionality reappeared. The GG revealed similar catastrophes in compositionality, though less frequently. In addition, the general trend revealed an increase in compositionality to an average level of $0.89 \pm 0.07$ at the end of the experiment.

Communicative accuracy for both game types also rose rapidly to values between 0.8 and 0.9 (Fig. 5 centre column). Although accuracy in the OG was not affected strongly from the fluctuations in compositionality, in the GG this was the case. Each time compositionality decreased, accuracy also decreased.

Coherence (graphs on the right) revealed more differences between the OG and the GG. Whereas coherence fluctuated largely between 0.4 and 0.8 in the OG without any clear in-

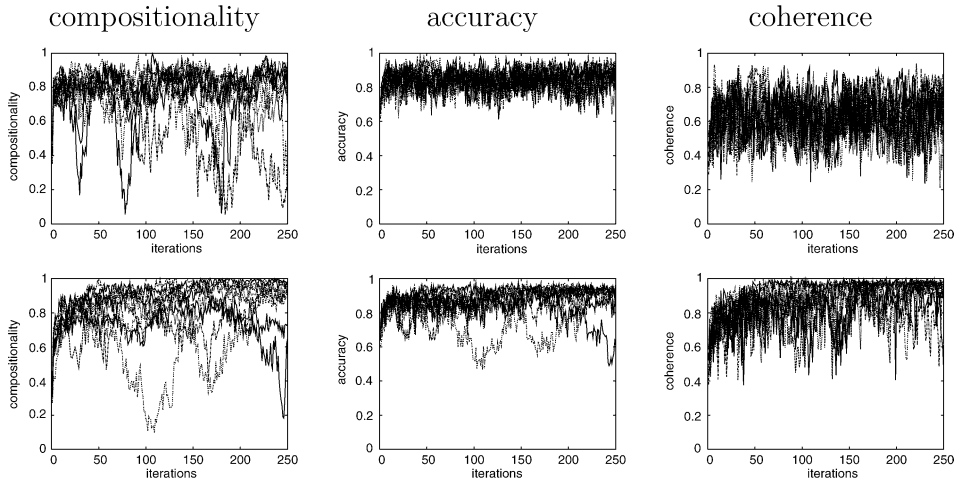


Fig. 5. The results of the baseline experiments of the observational games (top) and guessing games (bottom) show the evolution of compositionality (left), accuracy (centre) and coherence (right) on the y-axes. The x-axes show the iterations. Each line represents one trial.

Table 3
A fragment of a grammar that emerged in one of the baseline simulations of the guessing game

| | Adult | | Learner | |
|---|---|---|---|---|
| 1 | S → lfmia/[$0_r, 0_g, 0_b, .78_s, .67_s$] | 1.00 | S → lfmia/[$0_r, 0_g, 0_b, .67_s, .78_s$] | 0.92 |
| 2 | S → A/s B/rgb | 0.99 | S → A/s B/rgb | 1.00 |
| 3 | S → B/rgb A/s | 5.7E–5 | S → B/rgb A/s | 0.06 |
| 4 | S → C/r D/gbs | 1.2E–5 | S → C/r D/gbs | 0.11 |
| 5 | S → ibfdib/[$1_r, .75_g, .75_b, .5_s$] | 0.11 | | |
| 6 | A → ibf/[$.5_s$] | 1.00 | A → ibf/[$.5_s$] | 1.00 |
| 7 | A → fdjdgdmf/[$0_s$] | 1.00 | A → fdjdgdmf/[$0_s$] | 1.00 |
| 8 | A → ide/[$.63_s, .67_s$] | 0.98 | A → ide/[$.64_s, .67_s, .66_s$] | 0.94 |
| 9 | A → boncm/[$.59_s$] | 0.98 | A → boncm/[$.59_s, .58_s$] | 0.83 |
| 10 | A → b/[$.11_s$] | 0.77 | A → ggdkab/[$.59_s$] | 0.20 |
| 11 | B → dib/[$1_r, .75_g, .75_b, .69_g, .69_b$] | 0.98 | B → dib/[$1_r, .69_g, .69_b$] | 1.00 |
| 12 | B → fkm/[$1_r, 0_g, 0_b$] | 0.96 | B → fkm/[$1_r, 0_g, 0_b$] | 1.00 |
| 13 | B → m/[$1_r, 0_g, 1_b$] | 0.87 | B → m/[$1_r, 0_g, 1_b$] | 1.00 |
| 14 | S → dggdkab/[$0_r, 0_g, 1_b, .59_s$] | 0.20 | B → d/[$0_r, 0_g, 1_b$] | 0.02 |
| 15 | C → b/[$1_r$] | 0.08 | C → b/[$1_r, 0_r$] | 0.10 |
| 16 | S → bgl/[$0_r, 0_g, 1_b, .62_s$] | 0.11 | D → gl/[$0_g, 1_b, .61_s, .62_s$] | 0.11 |
| 17 | D → oncmm/[$0_g, 1_b, .58_s$] | 0.01 | D → oncmm/[$0_g, 1_b, .58_s$] | 0.01 |

creasing trend, coherence in the GG showed a steady increase toward values near 1 similar to the trend of compositionality, although the GG, too, revealed some large fluctuations.

Table 3 shows a fragment of the grammars of one adult and one learner at the end of a typical iteration in one simulation run of the GG. The fragment is rather small; typically learners acquired approximately 100–125 rules and adults ended up with more or less 125–150 rules. The adult of this particular example acquired 161 rules, the learner 110. In the

fragment, we can see that the adult and learner acquired both holistic rules (1, 5, 14, 16) and compositional rules (2–4). It is interesting to note that the rule weights of the learner's compositional rules are higher than those acquired by the adult. The table also shows how the adult's holistic rules 14 and 16 have become compositional in the learner's grammar. (Learner rules 3, 10 and 14 can decode adult rule 14, and the learner rules 4, 15 and 16 decode adult rule 16.) Rules concerning the expression "ibfdib" show a meaning shift (rules 2, 5, 6, 11). Note that the adult has two ways to construct "ibfdib". The grammar also contains some ambiguities in the form of polysemy (adult rules 1, 10, 11 and 15, and learner rules 1 and 15) and synonymy (learner rules 9 and 10). One could argue that rules 8, 9 and 16 are also polysemous, but the ambiguous categories are so close to each other that one also could argue that they constitute only one meaning.

Summarising, with a population of size 2, both the current OG and GG models show that compositionality can be achieved, even in the absence of a transmission bottleneck. The GG, however, appears more stable than the OG, but neither is completely stable.

### *4.2. Increasing population size*

In the second experiment series, the population size is increased from 2 to 6, which means that the population in these experiments consisted of 3 adults and 3 learners.

Fig. 6 shows the results of this experiment. For both the OG and GG, compositionality increased to a high level in the first few iterations. After that, compositionality soon decreased. For the OG, the final value of compositionality was on average $0.22 \pm 0.10$; for the GG this was $0.16 \pm 0.11$. This decrease was more drastic and instable for the OG than for the GG, where in some runs compositionality remained for a while or even recovered. It is striking to see that in the OG accuracy seems to benefit from the decrease in compositionality (it rose to $0.83 \pm 0.05$). The opposite is true for the GG, where accu-

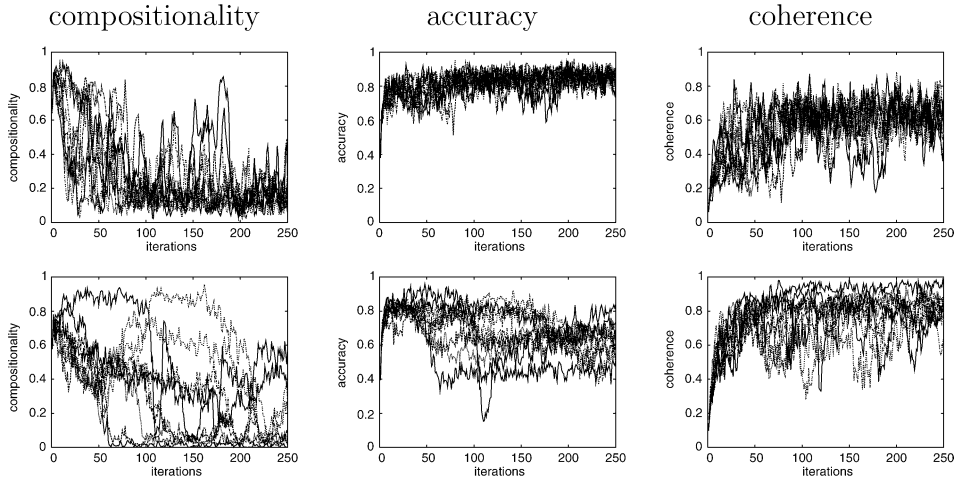


Fig. 6. The results of the experiments with a population size of 6. The figures show the results of the observational games (top) and guessing games (bottom) on compositionality (left), accuracy (centre) and coherence (right). Again, each line represents one simulation run.

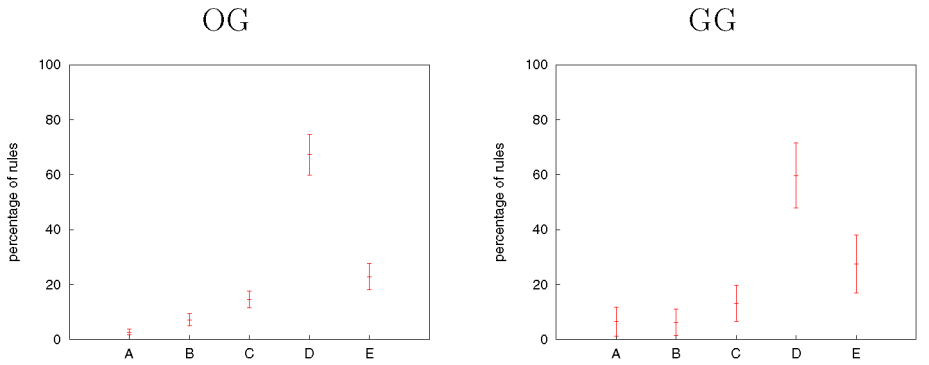


Fig. 7. The average percentage with which rules of types A–E have been used most frequently during one iteration for the baseline experiments of the observational game (OG) and the guessing game (GG). The types of rules are A: S → **rgb s**, B: S → **s rgb**, C: S → **r gbs**, D: holistic rules, and E: all other rules.

racy first decreased with the drop in compositionality, but then recovered slightly to a final value of $0.64 \pm 0.06$. In both cases, coherence increased gradually to various levels. In the OG, coherence reached a level of $0.57 \pm 0.09$, and in the GG a level of $0.82 \pm 0.07$ was reached. So, compositionality appears to have an antagonising effect on coherence: when compositionality decreases, coherence increases.

When looking at the grammars that evolved, we can measure various things. *Similarity*, for instance, measures the proportion with which the learner population has acquired the grammars of the adult population. At the end of the 250st iteration, similarity was found to be $0.71 \pm 0.05$ for the OG and $0.66 \pm 0.07$ for the GG. Throughout the evolutions, these values remained fairly constant.

Fig. 7 shows a distribution of different types of rules that emerged as *dominant* rules for an agent in an iteration. A rule is assumed to be dominant if it is used most frequently by an agent within one iteration. Since the most prominent structure in the environment combines colours with shapes, it is expected that if this structure is exploited most frequently, a dominance of rules of types A and B (i.e., rules that combine colour with shape) would be observed. In fact, the most dominant compositional rule found in both the OG and GG is one that combines the **r** component with the conceptual space covering **gbs**. Analysis of the **rgb** space for the used colours actually shows that for some colours the combination with equal values in the **r** dimension and different values in the **gbs** space occurs up to 5 times more often. However, as we would expect, given the low levels of compositionality, the most frequently used rules in all these experiments were holistic rules (type D).

### *4.3. Imposing a transmission bottleneck*

The third set of experiments is used to investigate what happens when a bottleneck on transmission is imposed. Fig. 8 shows the results of imposing a bottleneck of 50% on a population of size 2. Only compositionality is shown, which in both cases increase rapidly to a high level and remain there throughout the evolution. With $0.90 \pm 0.06$ compositionality in the GG is higher at the end than in the OG, which yielded $0.81 \pm 0.08$.

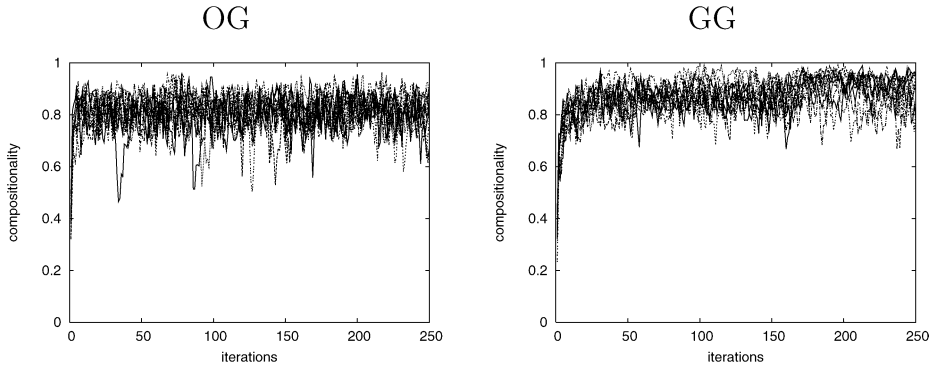


Fig. 8. Compositionality for a transmission bottleneck of 50% for a population size of 2. A transmission bottleneck of 50% means that all training phases considered only 60 out of 120 objects for inside the contexts. The results are shown for the OG (left) and GG (right).

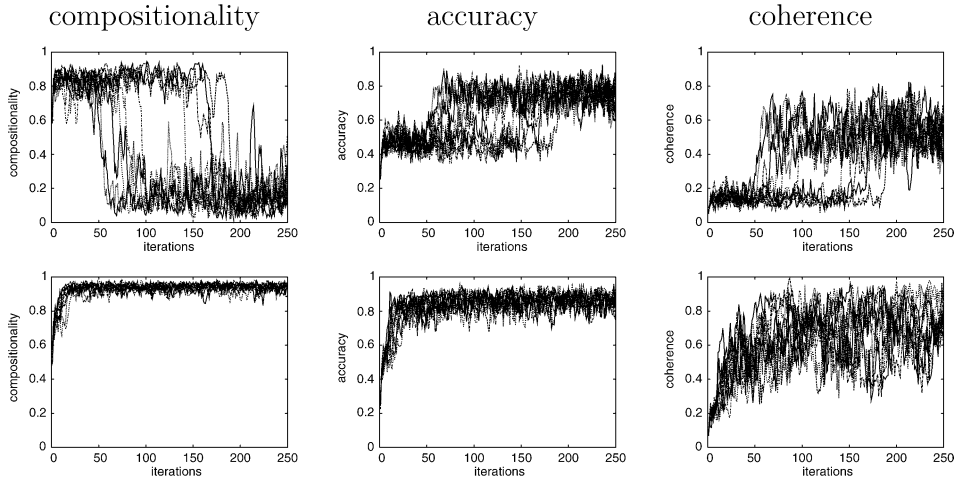


Fig. 9. The results of the experiments with a population size of 6 and a transmission bottleneck of 50%. The figures show the results of the observational games (top) and guessing games (bottom) on compositionality (left), accuracy (center) and coherence (right).

Fig. 9 shows the results of the simulations with a population of size 6 and a bottleneck of 50%. For the OG (top), compositionality tended to remain longer in the simulations than without a bottleneck, but at some point the languages colapsed into holistic systems within a few iterations (average of $0.23 \pm 0.10$ at the end). Whenever this happened accuracy and coherence jumped to a higher level. In the case of accuracy, this jump was from around 0.50 to 0.75; for coherence this was from around 0.16 to around 0.51. Similarity decreased slightly ($\pm 0.05$) with an overall average value of $0.71 \pm 0.02$ at the end.

The GG (bottom) increased very rapidly to a high and stable level of compositionality in all runs (an average of $0.94 \pm 0.02$ at the end). This was followed by accuracy, which rose rapidly to a level of around 0.8 after which it slowly kept on rising toward an average level

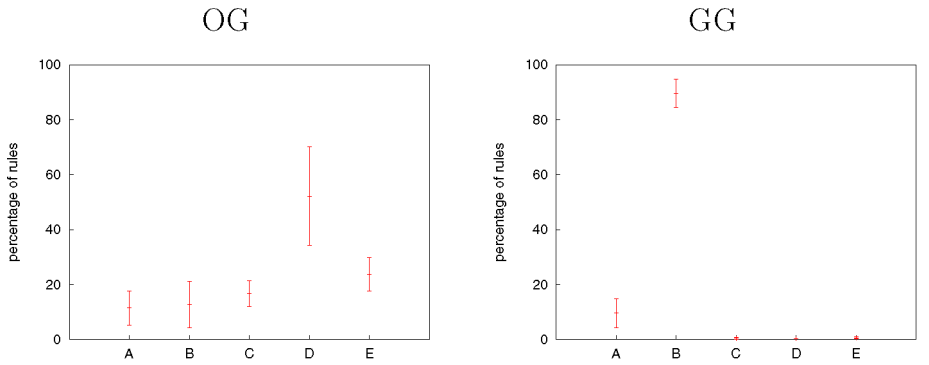


Fig. 10. The average percentage of rule types used in the experiments of the observational game (OG) and for the guessing game (GG) with a 50% bottleneck. The percentages relate to the rules of type A: S → **rgb s**, B: S → **s rgb**, C: S → **r gbs**, D: holistic rules, and E: rest.

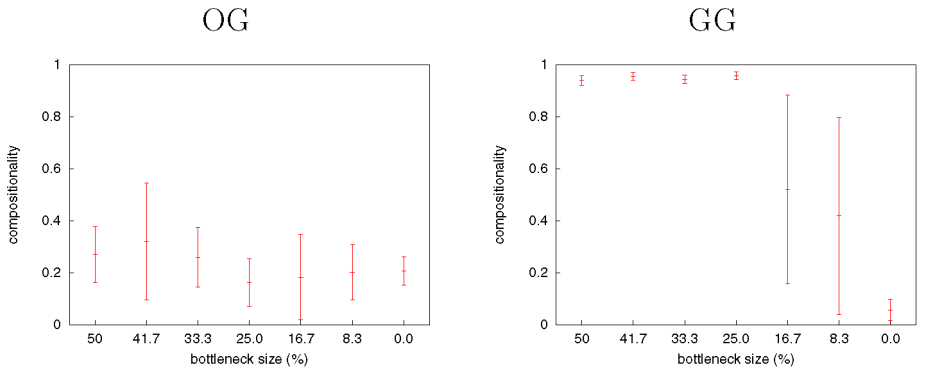


Fig. 11. Compositionality for various transmission bottlenecks for a population size of 6. A transmission bottleneck of 0% means that all objects were used in the training phase. The results are shown for the final (250st) iteration of the OG (left) and GG (right).

of $0.85 \pm 0.05$ in the end. Coherence yielded more fluctuations throughout the evolution, but its trend increased to an average value of $0.75 \pm 0.10$. After a sharp decrease of around 0.10 points in the first few iterations, similarity remained more or less constant with an overall average value of $0.65 \pm 0.02$.

Fig. 10 shows the distribution of dominant rules used throughout the experiments with a bottleneck of 50%. As could be expected, given the low level of compositionality, the distribution of the OG is much the same as without a bottleneck (Fig. 7), showing a peak in the holistic rules. As for the GG, nearly all dominant rules combine colours with shape, most of which prefer the word order with shape first (i.e., rules of type B: S → **s rgb**). This, therefore, nicely reflects the environment's structures.

Fig. 11 shows what happens to the level of compositionality at the end of the experiments when the bottleneck is varied from 0 to 50% (0% bottleneck means all objects are passed—i.e., there is no bottleneck). In all simulations with the OG, compositionality did not remain stable and eventually disappeared (as in the previous experiments, com-

positionality did emerge at first). In the GG, however, compositionality was high with a low standard deviation for bottlenecks of 25% or higher. When the bottleneck was 8.3% or 16.7%, a moderate level of compositionality was reached with a large standard deviation. (The large standard deviation indicates that sometimes compositionality was present, sometimes not.) Only when no bottleneck was imposed on the GG, compositionality did not remain stable.

## 5. Discussion

The previous section presented the results of the experiments done; in this section I will try to explain why the results are the way they are. After comparing the current model with some closely related work, I will discuss the rapid emergence of compositionality. Then I will discuss factors of instability with a primary focus on the differences between the observational and guessing games. The effect of increasing the population size is discussed in Section 5.4, after which I discuss the effect of the transmission bottleneck in Section 5.5.

### *5.1. Related work*

The current work is mostly related to the iterated learning model developed by Kirby [23,24]. Kirby's model starts with a population of size 2 (1 adult and 1 learner). Like in the current study, the adult produces utterances from which the learner learns its language. When the learner receives an utterance, it also observes its meaning. These meanings are predefined predicate argument structures of the form `p(x)` or `r(x,y)`. Apart from the selectionist learning mechanism and the instance-base used in the current paper, the learning mechanisms of both models are basically the same. The first major novelty in the current study is that the agents develop their own semantics and thus discover their own semantic structures, which reflect the combinatorial structure imposed by the environment. (This aspect is further elaborated in Section 5.2, but see also [60].) The second novelty is the application of the guessing game, including its selectionist learning mechanism. In this game, the learner only observes a context of possible objects from which it has to guess what the speaker's topic is. Kirby's model is more closely related to the observational game, in that the hearer/learner receives both the utterance and topic (albeit the exact meaning in Kirby's model and only the reference in this model). The third important novelty—with respect to Kirby's model—is the increased population size, but that has also been studied in [42]. Other related models using a 'Kirby' style implementation of the ILM are found in [6,41, 69]. All these studies predefine the semantics and have a population size of 2.

The learning mechanism is highly similar to Van Zaanen's alignment based learner [54], although he combines the alignment-based learner with probabilistic grammar inducers, similar to those described in [5]. Similar learning mechanisms have also been studied in [15,51]. Again, the main difference with these models—apart from its implementation in an evolutionary model—is that the semantics of the language develops from the agents perceptions of the world. Another difference is that alignments are derived from an instance-base that stores utterance-meaning pairs together with some new information, rather than combining the existing grammar with the new information. The reason for this

choice is that the instance-base contains more information about the most fruitful way to chunk up utterance-meaning pairs than could be done when only the grammar is used, unless the grammar is build using all possible parse-trees as is done in [5,54]. The problem with constructing all parse-trees is that in the current model it is unknown where the most optimal word boundaries are. Hence, there would be an extremely large number of rules that need to be stored. Alternatively, the model could rely only on the instance-base and use memory-based learning as in [14]. The problem there is its computational complexity. The current model has the advantage of relatively fast parsing (decoding), and only has to generalise from the instance-base in case of failures. (Note that Gong et al. [17] use a similar approach to study the emergence of compositionality, but they store an instance-base of utterance-meaning pairs in a buffer. If their buffer is full, they induce generalising rules and then empty the buffer.) Although the current implementation is still unrealistic, there is an increasing amount of evidence supporting the hypothesis of a dual storage of both exemplars and generalisations, e.g., [1,29].

Another related piece of work is by Steels and co-workers. First, the Talking Heads simulation is derived from Steels' Talking Heads experiment (TH) [49], which was an experiment studying the evolution of lexicons. Although slightly different, the input to the agents in the current simulation is very similar to the input that the TH received. In addition, where the TH used binary trees to represent categories, here categories are represented as 'prototypical' categories without a hierarchical layering. Furthermore, the TH allowed every combination of quality dimension to serve as a meaning, while here all 4 dimensions were required to form the meaning of the whole. (This was done in order to keep the current model simple; future work will allow more complex possibilities.) Recently, Steels has moved to a system in which robots (again embodied as cameras) develop a case-grammar from analysing scenes played in front of the camera [46]. This closely related work is far more complex than the current model, but seems to assume more. For instance, the lexicon and consequently the meanings are predefined—this omits the co-evolution of language and meaning, which may alter the outcome as we shall see shortly. In addition, the model does not have a population turnover, which is also true for Batali's [2] and Gong et al.'s [17] models. As the current model has shown, successful development of compositionality in the first iterations is no guarantee for success in later iterations. However, Steels' model relies heavily on the production of grammatical structures by the speakers and a recent study has shown that when the learners in the current model also act as speakers (i.e., they can speak to other learners or adults), compositionality remains stable, even in absence of a transmission bottleneck [61].

### *5.2. The rapid emergence of compositionality*

In most studies on the emergence of compositional structures using the iterated learning model, compositionality emerges at later stages of the evolution, e.g., [23,24]. So, why does compositionality emerge so rapidly in these experiments? To understand this rapid emergence, let us look at some differences between Kirby's model [23,24] and the current model. In Kirby's model, compositional rules can emerge, in which parts of an expression may have no semantic content. As a result, an uncontrolled growth of signal length can occur [42] and the chance of discovering alignments in the expressions is relatively low,

so the regularities in linguistic structures will emerge at a later stage. The current chunking algorithm is more restricted to split up expressions into two distinct substrings, each part of which must be semantically covered by a part of the whole meaning. In addition, this model has an inbuilt bias toward shorter strings. Following Zipf's law [68], the frequency distribution $f$ of words with string length $l$ is inversely proportional to this length, i.e., $f \propto l^{-\alpha}$, where $\alpha \approx 1$. Together with an alphabet size of 15 letters, the probability of finding aligning strings from the randomly generated words are high. A recent unpublished study, where the alphabet size was varied between 5 and 25, has revealed that when the alphabet size increases, the level of compositionality in the second iteration decreases proportionally. The level of compositionality at the end for a GG with 50% bottleneck, however was unaffected.

A second difference between Kirby's models and this one lies in the development of semantics. In Kirby's models, the semantics are predefined, but in the current model they are constructed during development. Recent analysis has shown that when agents have to discover a semantic structure from the Talking Heads environment, there are more possible combinations than when the structure is predefined in terms of a colour and shape space (i.e., compositional rules such as S → **r gbs**, S → **rg bs** etc. are also possible) [60]. Hence, as for the signal space with short signals and limited alphabet size, it is more likely to find a structure that can be exploited in compositional structures. Once compositional rules are used successfully, their weights are reinforced and more likely to be reused. As a compositional rule can be applicable in more situations than individual holistic rules, the rule weight of the compositional rule soon wins the competition from the rule weights of individual rules.

Another consequence of the co-evolution of meaning and grammar is that during development, the learner's representation of meanings differs from the adult's representation, as illustrated in Fig. 12. The figure shows a—for the illustration perfectly structured—language of an adult and an intermediate stage of a learner. The adult has 5 categories/

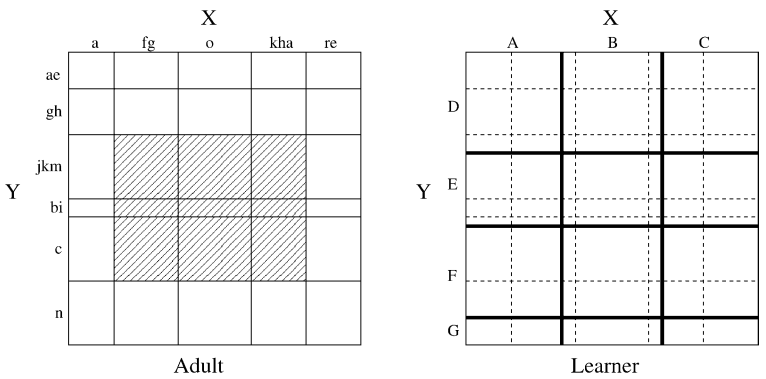


Fig. 12. Differences between an adult's and learner's language at a certain stage during the learner's development. The adult (left) has acquired an ontology that maps onto the world reliably, but the learner (right) has not yet acquired all these categories. In the adult's conceptual space, the categories that overlap with the learner's central category EB are highlighted. In the learner's conceptual space, the adult's categories are superposed as indicated by the dotted lines. See the text for further details.

words in the horizontal dimension and 6 categories/words in the vertical dimension. The learner has only acquired 3 categories in the horizontal dimension and 4 in the vertical dimension. Suppose that expressions are composed with the vertical dimension first, leading to expressions such as "aea", "aefg", "gha" etc. Furthermore, meanings can be constructed in a similar way, so the learner can have meanings DA, DB, EA, EB etc.

Around this stage of the development, the learner could have learnt the following rules:

| | | | |
|---|---|---|---|
| $S \rightarrow Y\ X$ | $\rho_1$ | | |
| $Y \rightarrow jkm/[D^{w1}, E^{w2}]$ | $\rho_2$ | $X \rightarrow fg/[A^{w6}, B^{w7}]$ | $\rho_5$ |
| $Y \rightarrow bi/[E^{w3}]$ | $\rho_3$ | $X \rightarrow o/[B^{w8}]$ | $\rho_6$ |
| $Y \rightarrow c/[E^{w4}, F^{w5}]$ | $\rho_4$ | $X \rightarrow kha/[B^{w9}, C^{w10}]$ | $\rho_7$ |

It is unlikely that the learner would have learnt all these rules in this way, but when it is exposed to the adult's language, it would undoubtedly have acquired some of these rules.

A learner with a limited set of categorical features, as in the early stages of development, is more likely to discover a semantic structure with respect to a syntactic structure than a learner who has a fully predefined, 'innate' set of categorical features, but no semantic structure. This effect, though small, is confirmed in [60], where it was shown that the initial level of compositionality was significantly lower when the categorical features were predefined. However, the same study also revealed that in the GG with a population of size 6, compositionality remained stable in the *absence* of a bottleneck. This latter phenomenon can be explained by the absence of a meaning shift (see below).

In summary, given the learning mechanisms for discovering and constructing compositional structures, the rapid emergence of compositionality depends to a large extent on the statistically high distribution of reoccurring structures in *both* the signal and meaning spaces. This statistical effect is increased by the ontogenetical development of categorical features. In addition, the selectionist reinforcement of rule weights further boosts the emergence of compositionality.

### *5.3. Factors of instability*

Given the seeming ease with which compositionality emerges, it is striking to see that it is not stable in all circumstances. In many cases, holistic communication systems take over after a while. This is especially surprising, given both the statistical nature of the data to be learnt and the learning mechanism with an inbuilt bias toward compositionality.[8]

As the instability most frequently occurs in simulations of the observational game, it is good to start investigating the instability factors with respect to this game. First it is important to realise that holistic communication systems have a higher level of accuracy and coherence in the OG. This indicates that—at least for the OG—a holistic system is easier to learn.

To see why this is the case, let us briefly return to the example illustrated in Fig. 12 and assume that the learner has acquired all signals, but not yet all categorical features

[8] Note that Kirby turns the surprise the other way around: he is surprised that compositionality emerges when there is a bottleneck, rather than being surprised that no compositionality emerges when there is no bottleneck [23].

that are used by the adult. Suppose now that at a later stage in development the learner acquired the same categorical features as the adult. Then the learner could have acquired three names ("fg", "o" and "kha") in relation to the category that the adult names "o" in the X dimension, which is a refinement of the earlier meaning B of the learner. Likewise, what the adult calls "bi" relates to the category that the learner could have associated with "jkm", "bi" and "c". Supposing there are 3 adults, the input to a learner is likely to be even more diverse, because the different agents may have slightly different categorical features themselves. So, different words could be spread even further through the conceptual space, shifting the meaning of a word, such as "o" to either side of the X dimension. Such shifts affect much of the conceptual space, making the language unstable.

In the OG, this is—in a certain way—allowed without affecting the accuracy too much, because in this game the hearer knows the speaker's topic. If the hearer has an association of the topic's meaning with the speaker's utterance, whether or not it is the hearer's preferred one, the game is considered successful. The weights are increased, while competing rules are laterally inhibited. But as there are already many associations/rules competing and different adults may use different utterances, the chances that another rule will take over are high. Consequently, a lot of unstable fluctuations are present in the competition, which makes the language hard to learn.

One the other hand, when the language is holistic, words are stored in more limited areas of the conceptual space. The chance of moving such a word through the conceptual space is much lower, causing fewer fluctuations, and is therefore easier to learn. When at some point, a less structured input is presented to a learner, this can have the consequence that compositional rules are less often successful, causing a decrease in their strength, making way for more holistic rules to win the competition. Once such a process starts, holistic rules become stronger and compositional rules weaker and the whole system turns around within a couple of iterations.

In the guessing games, this effect is less strong and only occurs when there is no bottleneck. This is because, in the GG, words are not allowed to move over the meaning space so easily. If this happens, the chance of failure in the game is high, because the hearers must *guess* the speaker's topic: the more possibilities there are, the more likely a failure. Failures do not come unnoticed, but are immediately punished by inhibiting such rules. In addition, when the right choices are made, these rules will be reinforced and competing rules inhibited, thus strengthening the disambiguation in the competition. As a consequence, the compositional rules are learnt properly—especially when there is a bottleneck. When there is no bottleneck, the holistic rules tend to enter the language more gradually. Though once a process is started, this can lead to a cascade of failures in compositional rules, which then make way for holistic rules.

In a way, this explains the difference between the OG and the GG in a similar way as was previously done in [55,62]. Because the OG has perfect joint attention, there is little pressure to disambiguate the language; in fact verbal communication becomes redundant as explained by the *signal redundancy paradox* [39]. This paradox states that if the agents have other means than language to establish joint attention, verbal communication becomes redundant. In the guessing game, there *is* a pressure to disambiguate ambiguities because the topic is not given beforehand, but must be guessed from the expression.

The instability in compositionality is also a sign that results achieved by studies that do not use a population turnover, might draw the wrong conclusion that once a compositional structure is learnt, this will persist in the language. This is important, because many studies base their results on experiments without a population turnover, such as [2,17,46].

### 5.4. The effect of population size

The first set of experiments in this paper used a population size of 2: 1 adult and 1 learner, which is the same as typically done in experiments using the ILM [6,7,23,24,41, 69]—but see [42]. The results of the current study differs in a large extent from the usual ILM studies in that the current experiments yielded an emergence of compositionality in the absence of a bottleneck, whereas this did not occur in the other experiments.

In addition, the current experiments showed that the model is well scalable to a larger population size, while this has been proved difficult with Kirby's ILM model [42]. Moreover, when the population size is increased, the characteristics of the other ILM studies reappear in that holistic languages emerge without a bottleneck and compositional ones with a bottleneck.

The reason why the experiments with a population size of 2 yielded relatively stable compositional systems in the absence of a bottleneck—as opposed to the experiments of population size 6—lies in the fact that a learner only learns from 1 adult. As a consequence, the probability that words become spread over the meaning space is less likely, and the bias towards compositional rules generally wins over the advantages of having holistic rules. However, the results also show that in some occasions, the system does become instable, which suggests that in those cases, the ambiguity of the compositional system becomes too large. Whether or not this is actually the case remains to be verified in future experiments.

When the population size is larger, the learning takes place with input from 3 different individuals. As explained above, this can lead to a meaning drift that makes the compositional system unstable when there is no bottleneck in the transmission. When there is a bottleneck, however, the GG yields highly stable compositional systems.

Compositionality in the current experiment emerged readily from the second iteration onward (at least when it remained stable). Each iteration was run for 6,000 games, which is equivalent to around 2,000 games for each learner to acquire the language of around 120 objects. Are these many language games really required for a stable communication system to emerge? Recent unpublished simulations have revealed that when a total of 2,000 language games (i.e., around 667 games per learner) are played, a stable and successful compositionality emerges for a guessing game in the presence of a bottleneck. How this will scale up in terms of population size, number of objects and complexity of the conceptual spaces (i.e., number of quality dimensions) remains to be seen. For vocabulary systems, a relation has been found between the time of convergence ($T$) in communicative success and the number of meanings $M$ and agents $N$ following $T \propto S \cdot N \cdot \ln(N)$ [21]. Gong et al. [17] have shown—in a similar simulation as the current one, but with a predefined semantics and without a population turnover—that compositional systems can emerge with a population of size 50, though many of their parameter settings (e.g., number of meanings) have not been specified. Preliminary studies on the further scalability of the guessing game model indicate that compositionality can remain stable in populations of

at least 100 agents, though stability is less frequently observed for increasing population sizes.

In addition, work is in progress to extend the current model in a very large scale simulation (1,000+ agents) on the evolution of language and culture in the recently started New Ties[9] project [63]. In this project many more objects and actions will become part of the agents' environment, which they can detect using more quality dimensions with fewer features, which—according to Brighton [6] can yield more stable compositional systems than systems of low dimension and large number of features (which is the case in the current experiment). A selection mechanism will then be required to select which quality dimensions will be used to form a sentence, rather than constraining sentences to cover all dimensions as in the current model. This selection criterion could be based on the ecological relevance and distinctiveness for the agents in a particular context; an aspect that the current model lacks and which may have a large impact on the development of language.

### *5.5. The influence of the bottleneck*

Imposing a bottleneck on the transmission of language leads to the emergence of compositionality in the GG. This confirms the results previously obtained in, e.g., [6,7,23, 24,41,69]. The emergence of compositionality under the influence of a bottleneck can be explained by noticing that the adults teach the learners about certain previously unseen objects. The use of compositional structures aids in doing so, as explained by the following example. Suppose an agent has acquired the rules

| **S**/**rgbs** → **A**/**rgb B**/**s** | |
|---|---|
| **A**/**rgb** → **blue**/$[0_r, 0_g, 1_b]$ | **B**/**s** → **square**/$[1_s]$ |
| **A**/**rgb** → **red**/$[1_r, 0_g, 0_b]$ | **B**/**s** → **triangle**/$[0_s]$ |

from observing a blue square, a blue triangle and a red triangle. This agent is able to communicate not only about these objects, but also about previously unseen red squares. As a result of using compositional rules successfully, these will be reinforced and hence be reselected and so on. The next generation will then observe more structure, allowing to preserve and build up even more structure.

Clearly the GG shows a high and stable level of compositionality under the presence of a bottleneck. The OG, however, does not reveal this when the population is larger than 2, although compositionality tends to remain longer in the population. Apparently, the lack of pressure for disambiguation in the OG (see above) is stronger than the pressure imposed by the bottleneck.

The size of the bottleneck does matter. It has been shown by Brighton that—under certain conditions—the stronger the bottleneck is, the higher the chance for stable compositional languages [6]. This result is confirmed by the GG, where for bottlenecks stronger than or equal to 25%, compositionality is very stable at more or less similar levels. When

[9] New Ties stands for: New Emerging World models through Individual, Evolutionary and Social learning. See http://www.new-ties.org.

the bottleneck is weaker, the stability is less secure (see the large standard deviations in Fig. 11), meaning that compositionality is sometimes stable, sometimes not.

In the current experiments, a *bottleneck on the transmission of language is imposed by the experimenter*. It has been shown, however, that when learners are allowed to act as speakers during their development, compositionality can emerge in the GG with a population size of 6 in the absence of the imposed transmission bottleneck [61]. (A more recent unpublished study indicates that this finding is even stronger when the population size is increased to 100 agents.) This is understood when realising that when learners start to speak, they are faced with a bottleneck, because due to their developmental stage, they may encounter previously unseen (or unheard) objects. If, however, they have learnt parts of the objects' meanings, they can exploit these by producing a compositional expression. This is interesting, as it may help to explain why children are so good at developing grammar, such as observed in normal situations [29], or perhaps at inventing grammar as in Nicaraguan Sign Language [36] and possibly creoles [35].

## 6. Conclusion

The model presented in this paper provides a promising framework for studying the emergence and evolution of compositional structures in language by exploiting both regularities found in the real world and (randomly generated) regularities found in linguistic surface structures. As both human languages and the world are highly structured, the model could be used profitable in the development of robotic platforms too.

From a scientific point of view, the experiments show that, in the model, compositional structures can emerge such that they reflect the structures of the world to a large extent. Factors influencing the emergence of compositionality include induction mechanisms, structures of the world, (randomly generated) structures in expressions and transmission bottlenecks. The latter confirms findings by [6,23,41,69] and can be used to argue that "the poverty of the stimulus solves the poverty of the stimulus" [69]. Moreover, the parallel development of syntax and semantics also appears to have a positive effect on the development of compositionality. The results also confirm earlier findings that the guessing game appears to provide a better strategy to evolve qualitatively more informative languages than the observational games do [55,62].

Current research focuses on scaling the experiments in terms of population sizes [63], and other parameters affecting the current model, such as the alphabet size. Future work will then scale the complexity of both the world and the agents in order to provide an environment for more complex languages to emerge. In addition, work is underway in order to extend the results of [61] regarding the potential ability to explain the grammatical creativity found in children.

## References

[1] R.H. Baayen, T. Dijkstra, R. Schreuder, Singulars and plurals in Dutch: Evidence for a parallel dual-route model, J. Memory and Language 37 (1997) 94–117.

[2] J. Batali, The negotiation and acquisition of recursive grammars as a result of competition among exemplars, in: E. Brisoe (Ed.), Linguistic Evolution through Language Acquisition: Formal and Computational Models, Cambridge University Press, Cambridge, 2002, pp. 111–172.
[3] T. Belpaeme, L. Steels, J. van Looveren, The construction and acquisition of visual categories, in: A. Birk, J. Demiris (Eds.), Learning Robots, Proceedings of the EWLR-6, in: Lecture Notes on Artificial Intelligence, vol. 1545, Springer, Berlin, 1998.
[4] P. Bloom, How Children Learn the Meanings of Words, MIT Press, Cambridge, MA, 2000.
[5] R. Bod, Beyond Grammar—An Experience-Based Theory of Language, CSLI Publications, Stanford, CA, 1998.
[6] H. Brighton, Compositional syntax from cultural transmission, Artificial Life 8 (1) (2002) 25–54.
[7] H. Brighton, S. Kirby, The survival of the smallest: Stability conditions for the cultural evolution of compositional language, in: J. Kelemen, P. Sosík (Eds.), Proceedings of the 6th European Conference on Artificial Life, ECAL 2001, in: Lecture Notes in Artificial Intelligence, vol. 2159, Springer, Berlin, 2001.
[8] E.J. Briscoe, Grammatical acquisition and linguistic selection, in: E. Briscoe (Ed.), Linguistic Evolution through Language Acquisition: Formal and Computational Models, Cambridge University Press, Cambridge, 2002, pp. 255–300.
[9] E.J. Briscoe (Ed.), Linguistic Evolution through Language Acquisition: Formal and Computational Models, Cambridge University Press, Cambridge, 2002.
[10] A. Cangelosi, D. Parisi, The emergence of "language" in an evolving population of neural networks, Connection Sci. 10 (1998) 83–93.
[11] A. Cangelosi, D. Parisi (Eds.), Simulating the Evolution of Language, Springer, London, 2002.
[12] N. Chomsky, Rules and representations, Behavioral Brain Sci. 3 (1980) 1–61.
[13] M.M. Chouinard, E.V. Clark, Adult reformulations of child errors as negative evidence, J. Child Language 30 (3) (2003) 637–669.
[14] W. Daelemans, A. van den Bosch, J. Zavrel, Forgetting exceptions is harmful in language learning, Machine Learning 34 (1999) 11–43.
[15] C.G. de Marcken, Unsupervised language acquisition, PhD thesis, Massachusetts Institute of Technology, 1996.
[16] P. Gärdenfors, Conceptual Spaces, Bradford Books, MIT Press, 2000.
[17] T. Gong, J. Ke, J.W. Minett, W.S.-Y. Wang, A computational framework to simulate the co-evolution of language and social structure, in: J. Pollack, M. Bedau, P. Husbands, T. Ikegami, R.A. Watson (Eds.), Artificial Life IX Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems, MIT Press, Cambridge, MA, 2004, pp. 214–219.
[18] T. Hashimoto, T. Ikegami, Emergence of net-grammar in communicating agents, Biosystems 38 (1996) 1–14.
[19] J.R. Hurford, Biological evolution of the saussurean sign as a component of the language acquisition device, Lingua 77 (2) (1989) 187–222.
[20] J.R. Hurford, Social transmission favours linguistic generalization, in: C. Knight, M. Studdert-Kennedy, J. Hurford (Eds.), The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form, Cambridge University Press, Cambridge, 2000, pp. 324–352.
[21] F. Kaplan, L'émergence d'un lexique dans une population d'agent autonomes, PhD thesis, Laboratoire d'infromatique de Paris 6, 2000.
[22] S. Kirby, Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners, in: C. Knight, M. Studdert-Kennedy, J.R. Hurford (Eds.), The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form, Cambridge University Press, Cambridge, 2000, pp. 303–323.
[23] S. Kirby, Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity, IEEE Trans. Evolutionary Comput. 5 (2) (2001) 102–110.
[24] S. Kirby, Learning, bottlenecks and the evolution of recursive syntax, in: T. Briscoe (Ed.), Linguistic Evolution through Language Acquisition: Formal and Computational Models, Cambridge University Press, Cambridge, 2002.
[25] S. Kirby, Natural language from artificial life, Artificial Life 8 (3) (2002).
[26] S. Kirby, J.R. Hurford, The emergence of linguistic structure: An overview of the iterated learning model, in: A. Cangelosi, D. Parisi (Eds.), Simulating the Evolution of Language, Springer, London, 2002, pp. 121–148.

[27] G. Lakoff, Women, Fire and Dangerous Things, The University of Chicago Press, Chicago, 1987.
[28] R.W. Langacker, Foundations of Cognitive Grammar, Stanford University Press, Stanford, CA, 1987.
[29] E. Lieven, H. Behrens, J. Speares, M. Tomasello, Early syntactic creativity: A usage-based approach, J. Child Language 30 (2) (2003) 333–370.
[30] B. MacLennan, Synthetic ethology: An approach to the study of communication, in: C.G. Langton, C. Taylor, J.D. Farmer (Eds.), Artificial Life II, in: SFI Studies in the Sciences of Complexity, vol. X, Addison-Wesley, Redwood City, CA, 1991.
[31] D. Marocco, A. Cangelosi, S. Nolfi, The emergence of communication in evolutionary robots, Philos. Trans. Math. Phys. Engrg. Sci. 361 (1811) (1996) 2397–2421.
[32] N. Neubauer, Emergence in a multiagent simulation of communicative behaviour, 2004.
[33] M. Oliphant, The dilemma of saussurean communication, Biosystems 1–2 (37) (1996) 31–38.
[34] S. Pinker, P. Bloom, Natural language and natural selection, Behavioral Brain Sci. 13 (1990) 707–789.
[35] G. Sankoff, S. Laberge, On the acquisition of native speakers by a language, Kivung 6 (1973) 32–47.
[36] A. Senghas, S. Kita, A. Özyürek, Children creating core properties of language: Evidence from an emerging sign language in Nicaragua, Science 305 (5691) (2004) 1779–1782.
[37] J.M. Siskind, A computational study of cross-situational techniques for learning word-to-meaning mappings, Cognition 61 (1996) 39–91.
[38] A.D.M. Smith, Intelligent meaning creation in a clumpy world helps communication, Artificial Life 9 (2) (2003) 559–574.
[39] A.D.M. Smith, Mutual exclusivity: Communicative success despite conceptual divergence, in: M. Tallerman (Ed.), Language Origins: Perspectives on Evolution, Oxford University Press, Oxford, 2005, pp. 372–388.
[40] K. Smith, The evolution of vocabulary, J. Theoret. Biol. 228 (1) (2004) 127–142.
[41] K. Smith, H. Brighton, S. Kirby, Complex systems in language evolution: The cultural emergence of compositional structure, Adv. Complex Syst. 6 (4) (2003) 537–558.
[42] K. Smith, J.R. Hurford, Language evolution in populations: Extending the iterated learning model, in: W. Banzhaf, T. Christaller, J. Ziegler, P. Dittrich, J.T. Kim (Eds.), Advances in Artificial Life: Proceedings of the 7th European Conference on Artificial Life, in: Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence, Springer, Heidelberg, 2003, pp. 507–516.
[43] L. Steels, Emergent adaptive lexicons, in: P. Maes (Ed.), From Animals to Animats 4: Proceedings of the Fourth International Conference On Simulating Adaptive Behavior, MIT Press, Cambridge MA, 1996.
[44] L. Steels, Perceptually grounded meaning creation, in: M. Tokoro (Ed.), Proceedings of the International Conference on Multi-Agent Systems, AAAI Press, Menlo Park, CA, 1996.
[45] L. Steels, The synthetic modeling of language origins, Evolution of Communication 1 (1) (1997) 1–34.
[46] L. Steels, Constructivist development of grounded construction grammars, in: W. Daelemans (Ed.), Proceedings Annual Meeting of Association for Computational Linguistics, 2004.
[47] L. Steels, T. Belpaeme, Coordinating perceptually grounded categories through language. A case study for colour, Behavioral Brain Sci. (2005), in press.
[48] L. Steels, F. Kaplan, Situated grounded word semantics, in: Proceedings of IJCAI-99, Stockholm, Sweden, Morgan Kaufmann, San Mateo, CA, 1999.
[49] L. Steels, F. Kaplan, A. McIntyre, J. Van Looveren, Crucial factors in the origins of word-meaning, in: A. Wray (Ed.), The Transition to Language, Oxford University Press, Oxford, UK, 2002.
[50] L. Steels, P. Vogt, Grounding adaptive language games in robotic agents, in: C. Husbands, I. Harvey (Eds.), Proceedings of the Fourth European Conference on Artificial Life, MIT Press, Cambridge, MA, 1997.
[51] A. Stolcke, Bayesian learning of probabilistic language models, PhD thesis, University of California at Berkeley, 1994.
[52] M. Tomasello, The Cultural Origins of Human Cognition, Harvard University Press, 1999.
[53] M. Tomasello, Do young children have adult syntactic competence?, Cognition 74 (2000) 209–253.
[54] M. van Zaanen, Alignment-based learning versus data-oriented parsing, in: R. Bod, K. Sima'an, R. Scha (Eds.), Data Oriented Parsing, Center for Study of Language and Information (CSLI) Publications, Stanford, CA, 2003, pp. 385–403.
[55] P. Vogt, Bootstrapping grounded symbols by minimal autonomous robots, Evolution of Communication 4 (1) (2000) 89–118.
[56] P. Vogt, The physical symbol grounding problem, Cognitive Syst. Res. 3 (3) (2002) 429–457.
[57] P. Vogt, Anchoring of semiotic symbols, Robotics and Autonomous Systems 43 (2) (2003) 109–120.

[58] P. Vogt, Iterated learning and grounding: From holistic to compositional languages, in: S. Kirby (ed.), Language Evolution and Computation, Proceedings of the Workshop at ESSLLI, 2003.
[59] P. Vogt, THSim v3.2: The Talking Heads simulation tool, in: W. Banzhaf, T. Christaller, P. Dittrich, J.T. Kim, J. Ziegler (Eds.), Advances in Artificial Life—Proceedings of the 7th European Conference on Artificial Life (ECAL), Springer, Berlin, 2003.
[60] P. Vogt, Meaning development versus predefined meanings in language evolution models, in: L. Pack Kaelbling, A. Saffiotti (Eds.), Proceedings of IJCAI-05, 2005.
[61] P. Vogt, On the acquisition and evolution of compositional languages, Adaptive Behavior (2005), in press.
[62] P. Vogt, H. Coumans, Investigating social interaction strategies for bootstrapping lexicon development, J. Artificial Societies and Social Simulation 6 (1) (2003), http://jasss.soc.surrey.ac.uk.
[63] P. Vogt, F. Divina, Language evolution in large populations of autonomous agents: Issues in scaling, in: Proceedings of AISB 2005: Socially Inspired Computing Joint Symposium, 2005.
[64] P. Vogt, A.D.M. Smith, Learning colour words is slow: A cross-situational learning account, Behavioral Brain Sci. (2005), in press.
[65] G.M. Werner, M.G. Dyer, Evolution and communication in artificial organisms, in: C.G. Langton, C. Taylor, J.D. Farmer (Eds.), Artificial Life II, in: SFI Studies in the Sciences of Complexity, vol. X, Addison-Wesley, Redwood City, CA, 1991.
[66] L. Wittgenstein, Philosophical Investigations, Basil Blackwell, Oxford, UK, 1958.
[67] A. Wray, Protolanguage as a holistic system for social interaction, Language and Communication 18 (1998) 47–67.
[68] G.K. Zipf, Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley, Cambridge, MA, 1949.
[69] W.H. Zuidema, How the poverty of the stimulus solves the poverty of the stimulus, in: S. Becker, S. Thrun, K. Obermayer (Eds.), Advances in Neural Information Processing Systems 15 (Proceedings of NIPS '02), MIT Press, Cambridge, MA, 2003.