Meaning development versus predefined meanings in language evolution models

Paul Vogt[1,2]
[1]Language Evolution and Computation, School of Philosophy, Psychology and Language Sciences
University of Edinburgh, 40 George Square, Edinburgh EH8 9LL, U.K.
[2]Induction of Linguistic Knowledge / Computational Linguistics
Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
paulv@ling.ed.ac.uk

To appear in the proceedings of IJCAI-05.

Abstract

This paper investigates the effect of predefining semantics in modelling the evolution of compositional languages versus allowing agents to develop these semantics in parallel with the development of language. The study is done using a multi-agent model of language evolution that is based on the Talking Heads experiment. The experiments show that when allowing a co-evolution of semantics with language, compositional languages develop faster than when the semantics are predefined, but compositionality appears more stable in the latter case. The paper concludes that conclusions drawn from simulations with predefined meanings, which most studies use, may need revision.

1 Introduction

The field of evolutionary linguistics is a rapidly growing field in contemporary cognitive science. Many studies are based on computational modelling, where the researchers typically study aspects of language evolution using models that integrate AI techniques such as multi-agent systems, evolutionary computation, machine learning, natural language processing and robotics. See [Cangelosi and Parisi, 2002] for an extensive overview.

One of the most prominent aspects of human language that is researched concerns the origins and evolution of grammatical structures, such as *compositionality*. Compositionality refers to representations (typically utterances in languages) in which the meaning of the whole is a function of the meaning of its parts. Studies into the origins and evolution of compositionality have yielded models that can successfully explain how compositionality may emerge. Most models have semantic structures built in, so the agents only have to acquire a mapping from signals to these meanings, together with their syntactic structures [Brighton, 2002; Kirby, 2002; Smith *et al.*, 2003]. Only few models have considered how compositional structures can arise through a co-evolution between syntax and semantics, where the semantics are grounded through interactions with the world and develop from scratch [Steels, 2004; Vogt, 2005].

Naturally, the two approaches yield different results. Whereas in [Brighton, 2002; Kirby, 2002; Smith *et al.*, 2003] it takes a number of generations until compositionality arises, in studies where the syntax co-develops with the semantics compositionality arises from the first generation [Steels, 2004; Vogt, 2005]. Why is this difference? In this paper, the answer is sought by starting off with an implementation of Vogt's model, which was based on Kirby's model, but without predefined semantics, and then comparing this with a model in which the semantics is predefined, as in Kirby's model.

The next section presents some background relating to Kirby's and Vogt's model. Section 3 presents the model which is our starting point. The results are presented in Section 4 and discussed in Section 5.

2 Background

In the context of this paper, compositionality is defined as a representation of which the meaning of the whole can be described as a function of the meaning of its parts. For instance, the meaning of the expression "red apple" is a function of the meaning of "red" and the meaning of "apple". As a consequence, it is possible to substitute one part with another to form a new meaning as in the expression "green apple". In contrast, there are holistic representations in which the meaning of the whole cannot be described as a function of the meaning of its parts. For instance, the expression "kick the bucket" in the meaning of `dying` is a holistic phrase.

It has been repeatedly shown that compositional structures can arise from initially holistic structures (i.e. structures with no compositionality) using the *iterated learning model* (ILM) [Brighton, 2002; Kirby, 2002; Smith *et al.*, 2003]. In the ILM the population at any time consists of adults and children.[1] The children learn from the linguistic behaviour of adults. After a learning episode (or *iteration*), the adults are removed from the population, the children becomes new adults and new children enter the population and the process repeats. Kirby and others have shown that, given an induction mechanism that can induce compositional structures, an initially holistic language can change into a compositional one after a number of iterations, provided the language is transmitted through a *bottleneck*.

The transmission bottleneck entails that children only observe a part of the expressible meanings of the language. Assuming the children are equipped with a learning mechanism

[1]In most ILMs there is only one adult and one child.

to discover and store compositional structures whenever possible, these structures tend to remain in the language because they allow an agent to communicate about previously unseen meanings. Suppose, for instance, you only have observed the expressions “ab”, “ad” and “cb” meaning `p(m)`, `q(m)` and `p(n)` resp.. If you have no way to discover a compositional structure, you would not be able to express the meaning `q(n)`. However, if you have the ability to acquire compositional structures such as `S -> A/x B/y`, where `A/m -> a` or `A/n -> c`, and `B/p(x) -> b` or `B/q(x) -> d`, you would be able to form the sentence “cd” to express the meaning `q(n)`. If an agent has acquired the language through a bottleneck, it may have to produce expressions about previously unseen meanings when it has become an adult. If there is no bottleneck, the children are expected to have learnt the entire language, so no compositionality is required and typically does not evolve.

In Kirby’s model, all agents (adults and children alike) are equipped with the same predefined predicate argument semantics. Naturally this is unrealistic, since it is widely acknowledged that human children are born without innate semantics. The question is therefore: to what extent does predefining the agents’ semantics influence the results of such simulations?

To investigate this question, Vogt [2005] implemented a simulation based on Kirby’s model, but without predefining the semantics. In Vogt’s model (described below), the semantics of individual agents develop in parallel with the language learning. This way, the semantics of adult agents differ from the children’s.

Vogt [2005] showed that, even without a bottleneck, relatively high levels of compositionality developed very early in the simulation. It was hypothesised that this rapid emergence of compositionality has to do with the statistical nature of the input given to the agents (both from the environment and signals). Furthermore, it was shown that under certain conditions, e.g., when a bottleneck on transmission was absent, this compositionality was unstable. In those cases, the languages gradually (though sometimes suddenly) transformed into holistic languages. One possible explanation for such a transition was based on the ontogenetical development of meaning.

The study in this paper investigates the effect of predefining the semantics as opposed to semantic development with respect to the rapid development of compositionality and to the stability of the compositional structures. More concrete, a number of simulations will be presented in which the model is gradually changed from the model reported by Vogt [2005] to a model with predefined semantics as reported in Kirby [2002]. The aim is to verify the two possible explanations mentioned.

3 The model

The model is implemented in Vogt’s Talking Heads simulator THSim [Vogt, 2003].[2] In this simulation, a population of agents can develop a language that allows them to communicate about a set of geometrical coloured objects that form the agents’ world. Language development is controlled by agents playing a series of *guessing games*, cf. [Steels, 1997]. The description of the model provided below lacks many details and motivations. Unfortunately this is unavoidable due to the lack of space in this paper. For further details, consult [Vogt, 2005].

[2] Available at http://www.ling.ed.ac.uk/˜paulv/thsim.html.

	action	result
1	sensing the environment	context
2	select topic	topic
3	discrimination game	meaning
4	decoding	expression
5	encoding	topic
6	evaluate success	feedback
7	induction	grammar

Table 1: A brief outline of the guessing game. Step 2 and 4 are performed only by the speaker, steps 5 and 7 only by the hearer, and all other steps by both agents.

The guessing game (GG) is briefly outlined in Table 1. The game is played by two agents: a speaker and a hearer. Both agents sense the situation (or *context*) of the game. The context C consists of a given number of geometrical coloured objects. For each object o_j, the agents extract a feature vector $\mathbf{f}_j = \{f_r, f_g, f_b, f_s\}$, where f_i are features in a specified *quality dimension* [Gärdenfors, 2000]. These dimensions relate to a sensed quality, which in the current implementation is one of the components of the rgb colour space (f_r, f_g, f_b) or a shape feature f_s indicating the shape of the object.

The speaker selects one object from the context as the *topic* o_t of the game (step 2, Table 1) and plays a *discrimination game* [Steels, 1997] to find a category that distinguishes the topic o_t from the other objects in the context $o_j \in C\backslash\{o_t\}$ (step 3). To this aim, each individual agent constructs an ontology containing *categorical features* (CF), which are points c_{ik} in some quality dimension i. Each feature f_i is categorised with that categorical feature c_{ik} such that the distance $|f_i - c_{ik}|$ is smallest.

Combining the CFs of each quality dimension constructs a category $\mathbf{c}_j$ for object o_j. The category $\mathbf{c}_j$ is a point in a n-dimensional space. If the category for the topic o_t is different from the categories of all other objects in the context, the DG succeeds. If no distinctive category is found, then each feature f_i of the topic’s feature vector $\mathbf{f}_t$ is used as an exemplar to form new CFs c_{ik}, which are added to the ontology. This way, the ontology is built up from scratch.

The distinctive category resulting from the DG serves as the ‘whole meaning’ of o_t, where the whole meaning can be considered as a holistic semantic category. Semantic categories are represented by *conceptual spaces* [Gärdenfors, 2000]. A conceptual space is spanned by n quality dimensions and contain n-dimensional categories $\mathbf{c}_j$ formed from n CFs c_i. The holistic semantic category is spanned by all 4 quality dimensions. Other semantic categories include any possible configuration of quality dimensions, such as a colour conceptual space (spanned by the `rgb` qualities), a 1-dimensional shape space and a ‘redgreen’ space spanned by the `r` and `g` components of the `rgb` space. The whole

R_1	`S -> greycircle/`	$[.5_r, .5_g, .5_b, .6_s]$
R_2	`S -> A/rgb B/s`	
R_3	`A -> red/`	$[1_r, 0_g, 0_b]$
R_4	`B -> square/`	$[1_s]$
R_5	`B -> triangle/`	$[0_s, .11_s]$

Table 2: A constructed example grammar. In the example S, A and B are non-terminals. R_1 is a holistic rule that rewrites to an expression `greycircle` with meaning $[.5_r, .5_g, .5_b, .6_s]$. R_2 is a compositional rule that rewrites to non-terminals A and B that form linguistic categories associated with the semantic categories indicated by the quality dimensions `rgb` (for the colour space) and `s` (for the shape space). Each non-terminal rewrites to another expression and meaning. (Note that an expression can be associated with more than one meaning as in R_5.)

meaning can thus be formed by a holistic category, of by a combination of more lower dimensional categories. There are two constraints in the current model: (1) All quality dimensions should be used exactly once. (2) Only combinations of 2 semantic categories are allowed. During their development, agents learn (or *discover*) which semantic categories are useful to form a compositional language. (In one experiment, though, these are predefined.)

A composition $\oplus R_i$ may be of only one holistic rule (e.g., $\oplus R_i = R_1$ in Table 2), or of several rules, such as $\oplus R_i = R_2 \circ R_3 \circ R_4$, which combines a colour with shape (note that the operator $\circ$ is used to connect terminal rules (R_3 and R_4) with compositional rules such as R_2). Each rule R_i has a rule weight ρ_i and – for the terminal rules – each possible meaning of a rule has the weight weight μ_j, both indicating the effectiveness of a rule and meaning.[3] These weights are used to compute a score s_i indicating the effectiveness of the rule based on past experiences. If the rule R_i is a terminal rule, $s_i = \rho_i \mu_j$; if R_i is a compositional rule (such as R_2), $s_i = \rho_i$. Given the composition $\oplus R_i$, one can calculate the composition score $\sigma(\oplus R_i) = \prod_{\oplus R_i} s_i$. When a rule/meaning pair is used successfully in the guessing game, the weights ρ_i and μ_j are increased, while the weights of competing rules/meanings are inhibited. The score $\sigma(\oplus R_i)$ is used to select among competing compositions.

When the speaker obtained a distinctive category from the DG, it tries to *decode* an expression (step 4, Table 1). To this aim, the speaker searches its grammar G for compositions of rules that match the distinctive category (or meaning). So, the meaning $[.5_r, .5_g, .5_b, .6_s]$ can be decoded into the expression `greycircle` using R_1 and meaning $[1_r, 0_g, 0_b, 1_s]$ into `redsquare` using $R_2 \circ R_3 \circ R_4$. (Note that elements such as 1_r and 1_s are CFs with value 1 and dimension r and s respectively.) If there is more than one composition that can decode a meaning, the composition with the highest score $\sigma(\oplus R_i)$ is selected. This implements a selection mechanism favouring effective compositions.

If there is no composition that decodes the meaning, a new rule is constructed that either decodes a part of the meaning, or that decodes the meaning holistically. The former case is true if the speaker wanted to decode, for instance, meaning $[1_r, 0_g, 0_b, .6_s]$ using the grammar of Table 2. The part $[1_r, 0_g, 0_b]$ can be decoded using composition $R_2 \circ R_3 \circ ?x$, where $?x$ is the undecoded part. A new rule of the form `B -> circle/`$[.6_s]$ is then constructed to fill in $?x$. In the model, expressions are constructed at random from a limited alphabet Σ of size $|\Sigma| = 15$ and with length l, where $2 \leq l \leq 8$, with a frequency distribution $p(l) \propto \frac{1}{l}$ to implement a bias toward short strings.

The decoded expression is uttered to the hearer, which tries to *encode* this expression (step 5). The hearer does not know which object $o_i \in C$ is the topic, and its aim is to guess this using the verbal hint received. To do this, the hearer first plays a DG for each individual object $o_i \in C$, thus resulting in a set of possible meanings for the expression. Then the hearer searches its grammar for compositions that can parse the expression and of which the meaning matches one of the possible meanings. If there are more than one, the hearer selects the one with highest score $\sigma(\oplus R_i)$ and guesses that the object of the corresponding meaning is the topic.

This information is passed back to the speaker, who verifies if the guess is correct and provides the hearer with feedback regarding to the outcome (step 6). If the game succeeds, the weights of used rules are increased, while competing ones are inhibited using the equation $w = \eta w + (1 - \eta)\chi$, where w is a weight (ρ_i or μ_j), and $\chi = 1$ if the weight is increased and $\chi = 0$ if it is inhibited. (Note that a rule is competing if it can decode an expression for the topic or encode the expression received. Meanings compete when they belong to the same used rule, but are not the ones used, see, e.g., R_5.) If the game fails, the speaker provides the hearer with the correct topic and the hearer will then try to induce new knowledge (step 7).

Induction proceeds in up to three of the following steps (if one step succeeds, induction stops):

1. **Exploitation.** If there is a composition that partially decodes the expression, the remaining gaps are filled. For instance, if the hearer hears "redcircle" and the topic (indicated by the speaker) has distinctive category $[1_r, 0_g, 0_b, .6_s]$, then it can decode the expression partially using composition $R_3 \circ R_4 \circ ?x$, covering meaning $[1_r, 0_g, 0_b]$. In that case, the hearer will add the rule `B -> circle/`$[.6_s]$ to its grammar.

2. **Chunking.** If exploitation fails, the hearer will investigate whether it can chunk the expression-meaning pair based on stored (holistic) expression-meaning pairs received in the past. (These pairs are stored in an instance base.) This is done by looking for alignments in the received expression with stored expressions and alignments in the distinctive category with stored meanings. If more chunks are possible, the chunk that has occurred most frequently or that has the largest common string is actually pursued.

   For instance, if the instance base contains an entry with expression-meaning pair "pinkcircle"-$[1_r, .7_g, .7_b, .6_s]$ and the hearer tries to induce knowledge from receiving the expression-meaning pair "greencircle"-

[3] The actual implementation of μ_j is more complex, but for the purpose of this paper, the current explanation suffices.

$[0_r, 1_g, 0_b, .6_s]$, the following alignment is present: "green<u>circle</u>"-$[0_r, 1_g, 0_b, \underline{.6_s}]$.[4] Based on this alignment the following new rules are added: `S -> C/rgb D/s`, `C -> pink/`$[1_r, .7_g, .7_b]$, `C -> green/`$[0_r, 1_g, 0_b]$ and `D -> circle/`$[.6_s]$.

3. **Incorporation.** If chunking fails too, the hearer will incorporate the expression-meaning pair holistically. For instance, if the expression-meaning pair is: "wateva"-$[0_r, .2_g, .2_b, .7_s]$, the rule `S -> wateva/`$[0_r, .2_g, .2_b, .7_s]$ is added to the grammar.

At the end of the induction step, a post-operation called *generalise and merge* is performed to generalise even further and to eliminate redundancies. Redundancies can occur when new rules are created that are basically similar to already existing ones. The above example of chunking, for instance, yielded the construction of the rule `S -> C/rgb D/s` with additional filler rules. This rule, however, is in essence the same as rule R_3 from Table 2. Therefore, this new rule is removed (*merged*) and the non-terminals of the additional filler rules are replaced with the letters `A` and `B`. In addition, given a newly added rule `B -> circle/`$[.6_s]$ and the rules of Table 2, rule R_1 can be chunked (*generalised*) as well, leading to the new rule `A -> grey/`$[.5_r, .5_g, .5_b]$. (Note that the old rules such as R_1 remain in the grammar.)

The language game model is integrated with the iterated learning model. In this model, the population contains N adults and N children. Whenever a GG is played, the speaker is selected from the adult population and the hearer from the child population. After a given number of guessing games, all adults are replaced by the children, and new children enter the population. Such an iteration then repeats.

4 Results

A number of experiments were done with the above model, three of which are reported here. Each condition was carried out both without and with a transmission bottleneck. If a transmission bottleneck was imposed, a set of 60 objects were selected at random from the world of 120 objects at the start of each iteration. These 60 objects were then the objects about which the agents could communicate during that iteration.

Each experiment contained a population of 3 adults and 3 children. The experiments were run for 250 iterations of 6000 guessing games each. At the end of an iteration, the population was tested before the adults were replaced by the children and new children entered the population. In each test phase, 200 contexts were constructed. In every context one object was chosen as topic, and all agents then tried to encode an expression, which in turn all other agents tried to decode. From this phase, various measures were calculated of which compositionality is the measure reported in this paper. Compositionality is calculated as the ratio between the number of compositional rules used (both encoded and decoded) and the total number of utterances produced and interpreted. The testing phase always used the entire world of 120 objects.

[4] Alignments are only allowed if they appear at the start or the end of an expression, and if they are connected.

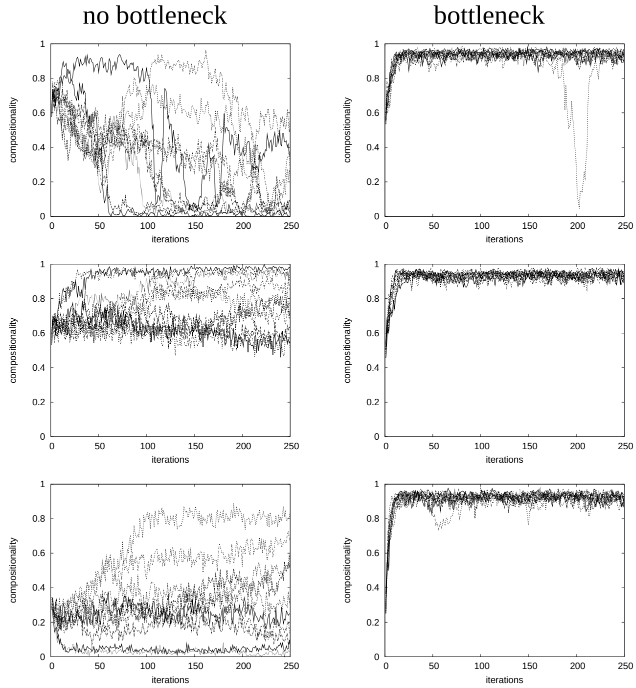

Figure 1: Compositionality as a function of iterations for experiments 1 (top), 2 (middle) and 3 (bottom). All experiments were done both without (left) and with (right) a transmission bottleneck. Each line represents one simulation run.

	no bottleneck		bottleneck	
	2	250	2	250
1	0.74±0.01	0.16±0.11	0.69±0.04	0.93±0.01
2	0.68±0.01	0.76±0.14	0.64±0.04	0.93±0.02
3	0.30±0.03	0.35±0.26	0.48±0.03	0.91±0.02

Table 3: The results of the first three experiments, both without and with bottleneck. For each condition compositionality is given for the end of the 2nd iteration and the end of the 250st iteration.

Experiment 1 The first experiment was done with exactly the same model as described in the previous section. Here all meanings developed from scratch in parallel to the development of the grammar.

Experiment 2 In experiment two, all agents were given a repertoire of categorical features. With these CFs, the agents could form categories right from the start, but they did not have any idea how to combine the different dimensions to form conceptual spaces as the basis of the semantic structures.

Experiment 3 In the third experiment, the semantic structure was predefined. All agents were given CFs as in experiment 2 and an additional constraint that semantic categories (i.e. conceptual spaces) were either colours, shapes or the whole (holistic) meaning.

Figure 1 and Table 3 show the results of these three experiments both without (left) and with (right) a transmission bottleneck. The development of compositionality in experiment 1 is shown in the top graphs. As can be seen, in both cases compositionality emerged to a rather high level in the first few iterations (0.74 ± 0.01 without and 0.69 ± 0.04 with a bottleneck, see Table 3, row 1). After that, compositionality tended to be instable and eventually died away in favour of holistic languages if no bottleneck was present (this was confirmed in simulations run for 500 iterations). With a transmission bottleneck, compositionality rapidly kept rising to a stable value of around 0.93. Only one simulation showed a sudden dramatic decrease in compositionality around iteration 200, but soon recovered to its previous level.

When the agents have predefined CFs, but had to learn the semantic structures (experiment 2, middle graphs), the initial levels of compositionality were slightly lower than in experiment 1. In the case where there was no bottleneck, compositionality remained in most cases at the same level, and in some cases even increased to levels near 1. When there was a bottleneck, compositionality rose in the same way to similar levels as in experiment 1.

In the case where the semantics, including their structures, were predefined (experiment 3, bottom graphs), the results were quite different from the other two experiments. In both cases, the initial levels were substantially lower, although – in contrast to the other experiments – compositionality was lower without a bottleneck (0.30 ± 0.03) than with a bottleneck (0.48 ± 0.03). When no bottleneck was imposed, compositionality increased to 0.8 in one simulation run, it rose in a few other cases, but often it remained at the same level or decreased. When a bottleneck was imposed, the compositionality revealed a similar increase as in the two other experiments with a bottleneck.

5 Discussion

The experiments in this paper investigate the hypothesis that the statistical nature of the semantic structure can explain the rapid development of compositionality and the hypothesis that the instability of compositionality in some circumstances had to do with the ontogenetical development of meaning. The experiments further investigate the effect of imposing a bottleneck on the transmission of language on the development of compositionality.

All experiments reveal that when a bottleneck is imposed, compositionality develops rapidly to a high and stable degree, thus confirming the results achieved earlier by Kirby and others [Brighton, 2002; Kirby, 2002; Smith *et al.*, 2003]. When no bottleneck is imposed, the behaviour of the different experiments is more different from each other, which will be discussed in the remainder of this section.

There are two types of observations: First, when no semantic *structure* is imposed (experiments 1 and 2), high levels of compositionality are achieved within two iterations. Second, when no categorical features are predefined (experiment 1), compositionality is unstable, whereas in other cases, compositionality is either stable (experiment 2) or may emerge at a later stage (experiment 3).

Let me start discussing the first observation. The problems the agents learn is to discover regularities in the semantic space (which reflects the feature space of the objects) in combination with regularities in the signal space. It was hypothesised that the rapid development of compositionality in experiment 1 (and 2) was caused by the statistical properties of the semantic space on one hand and the signal space on the other [Vogt, 2005]. Because signals tend to be short and are constructed from a limited alphabet, the likelihood of finding alignments in this space is high. Since this aspect is not varied, all experiments had the same likelihood.

X	Y	$P_1 \cdot P_2$
r	gbs	0.297
g	rbs	0.200
b	rgs	0.256
rg	bs	0.117
rb	gs	0.144
gb	rs	0.117
rgb	s	0.075

Table 4: Probability of finding co-occurring structures in dimension X in two different games, assuming the values in Y differ. Assuming that x_i can have any possible value in the space of X (indicated by $x_i \in X$) and y_i any value in the space of Y, then $P_1 = P(x_1 \in X \wedge y_1 \in Y)$ and $P_2 = P(x_2 = x_1 \wedge y_2 \in Y \backslash \{y_2\}))$. The values are based on the distribution of feature values, which are not presented in this text.

The chances for finding regularities in the semantic space, however, is much higher when no structure is predefined. In experiment 3, the structure for combining the `rgb` space with the shape `s` space was imposed. Given there are 12 colours and 10 shapes, there are $12 \times 10 = 120$ different ways to form semantic structures of these 120 objects. However, when no structure is imposed, the combinations `r-gbs`, `g-rbs`, `b-rgs`, `rg-bs`, `rb-gs` and `bg-rs` are possible ways to combine the 4 quality dimensions as well. Since, for instance, in the `r` dimension there are 40 different objects (4 colours and 10 shapes) that can have feature value $f_r = 0$. In a rough estimation, the probability P_1 of finding a one of these objects in game is proportional to $\frac{40}{120}$. The probability P_2 of finding *another* object with $f_r = 0$ in another game is $P_2 \propto \frac{39}{120}$. So, the probability of finding a co-occurring structure in the `r` dimension with $f_r = 0$ in two consecutive games is: $P_1 \cdot P_2 \propto 0.108$. Now, if we look at objects of the same colour, e.g., blue (i.e. with $f_r = 0$, $f_b = 1$ and $f_g = 0$), then the chance of finding a blue object in the first game is $P_1 \propto \frac{10}{120}$ and a different blue object in the second game is $P_2 \propto \frac{9}{120}$, so finding this co-occurrence is $P_1 \cdot P_2 \propto 0.006$.

Table 4 shows the co-occurrence probabilities for all objects involving similar values in one or more dimensions of the colour space. The values shown are the probabilities summed over all possible values in a conceptual space. As the table shows, when considering the entire colour space `rgb`, the co-occurrence probability is much lower than when searching for a co-occurring structure in the `r` space. So clearly, when an agent is able to construct any possible combination of semantic categories as in experiment 1, the chances of finding co-occurring structures is much higher than when it is restricted only to using `rgb-s` as in experiment 3. Hence, compositionality in experiment 1 emerges faster than in experiment 3.

Interestingly, when there is a bottleneck on the transmission of language and stable compositional systems emerge, the most dominant rules (i.e. the type of rules that were used most frequently) combine the colour space `rgb` and the shape space `s`. When there is no bottleneck, the most frequently used compositional rule combined `r` with `gbs`. This can be understood by realising that when half of the objects are discarded, the co-occurrence probability in the `rgb-s` structure is more or less maintained, while the other probabilities are affected differently in each iteration. This is because the distribution of colours and shapes is less skewed than the distributions in the other dimensions. (Recall that each iteration a different set of objects is selected to learn from.) As a result, the `rgb-s` structure can be learnt most reliably.

The second observation that compositionality is unstable in the absence of bottlenecks when categorical features are not predefined can be explained by realising the consequences of the gradual development of CFs. The foremost difference is that at the start of an iteration adults have a well developed set of CFs, while the children have none. During development, the children gradually acquire a similar set of CFs, but initially these CFs are overgeneralisations of the final categories. Nevertheless, the children may – in certain situations – categorise some objects distinctively, thus successfully finishing a discrimination game.

As a consequence, these children can successfully learn the words associated with these categories. However, since the CFs are overgeneralisations of the adults' CFs, the children's CFs may be associated with more than one word. This is even more the case when realising that different adults can have different words to name a CF. When the children acquire more fine grained CFs, the words may become associated with different CFs than was the case for the adults. This way, the words can drift through the conceptual space. When the language is compositional, the movement in one semantic category affects a larger part of the language than it would when the language is holistic. Hence, holistic languages are more stable and thus easier to learn. When the CFs are predefined, there is less or no reason for overgeneralisations. This, in turn, allows children to acquire any compositional structure of their adults more reliably, which – in part – explains why compositionality is more stable (experiment 2) or even emerges (experiment 3). The results of experiment 3 are very similar to those reported in [Smith *et al.*, 2003], while the results of experiment 2 is highly dissimilar. It is yet unclear whether the predefined semantics of Smith et al. are more similar to the ones used in experiment 2 or 3, though the results suggest that the semantics are more similar to the one used in experiment 3.

Concluding, both hypotheses tested in this paper are confirmed. In addition, the major findings of [Brighton, 2002; Kirby, 2002; Smith *et al.*, 2003] that compositionality emerges under the pressure of a transmission bottleneck are confirmed in all tested conditions. Hence, the assumption they made on predefining semantics to simplify their models appears valid. However, predefining the meanings – either with or without structure – can affect the results achieved in simulations of language evolution to a great extent, so one should be very careful in interpreting results achieved with such ungrounded models. This does not mean that the current grounded model on language evolution is sufficiently realistic to make hard claims regarding the evolution of language. However, the model is more realistic than ungrounded models in that in this model children do not have the same full-fletched semantic structures as adults.

Acknowledgements

This work is sponsored by a EC Marie Curie Fellowship and a VENI grant provided by the Netherlands Organisation for Scientific Research (NWO). The author thanks Andrew and Kenny Smith and three anonymous reviewers for comments on earlier versions of this manuscript.

References

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

[Cangelosi and Parisi, 2002] A. Cangelosi and D. Parisi, editors. *Simulating the Evolution of Language*. Springer, London, 2002.

[Gärdenfors, 2000] P. Gärdenfors. *Conceptual Spaces*. Bradford Books, MIT Press, 2000.

[Kirby, 2002] S. Kirby. Learning, bottlenecks and the evolution of recursive syntax. In Ted Briscoe, editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press, 2002.

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

[Steels, 1997] L. Steels. The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1–34, 1997.

[Steels, 2004] L. Steels. Constructivist development of grounded construction grammars. In W. Daelemans, editor, *Proceedings Annual Meeting of Association for Computational Linguistics*, 2004.

[Vogt, 2003] P. Vogt. THSim v3.2: The Talking Heads simulation tool. In W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, and J. Ziegler, editors, *Advances in Artificial Life - Proceedings of the 7th European Conference on Artificial Life (ECAL)*. Springer Verlag Berlin, Heidelberg, 2003.

[Vogt, 2005] P. Vogt. The emergence of compositional structures in perceptually grounded language games. Submitted for publication, 2005.