Evolution of an Optimal Lexicon under Constraints from Embodiment

Willem Zuidema
Language Evolution and Computation Research Unit
School of Philosophy, Psychology and Language Sciences
Institute for Cell, Animal and Population Biology
University of Edinburgh
40 George Square
Edinburgh EH8 9LL
United Kingdom
jelle@ling.ed.ac.uk

Gert Westermann
Centre for Brain and Cognitive Development
Birkbeck College
University of London
Malet Street
London WC1E 7HX
United Kingdom
g.westermann@bbk.ac.uk

Abstract Research in language evolution is concerned with the question of how complex linguistic structures can emerge from the interactions between many communicating individuals. Thus it complements psycholinguistics, which investigates the processes involved in individual adult language processing, and child language development studies, which investigate how children learn a given (fixed) language. We focus on the framework of *language games* and argue that they offer a fresh and formal perspective on many current debates in cognitive science, including those on the synchronic-versus-diachronic perspective on language, the embodiment and situatedness of language and cognition, and the self-organization of linguistic patterns. We present a measure for the quality of a lexicon in a population, and derive four characteristics of the optimal lexicon: specificity, coherence, distinctiveness, and regularity. We present a model of lexical dynamics that shows the spontaneous emergence of these characteristics in a distributed population of individuals that incorporate embodiment constraints. Finally, we discuss how research in cognitive science could contribute to improving existing language game models.

Keywords
Language, lexicon, evolution, self-organization, embodiment

1 Introduction

There exists a long tradition of formulating and studying formal models of language processing and language learning. These models have generally focused on the linguistic competence of a single individual. They have proven to be appealing because such formalisms offer precision and clarity, have led to successful technology, and have allowed for extensive theoretical research to complement empirical work.

However, these competence models have abstracted away many arguably crucial characteristics of language. These abstractions are viewed with growing uneasiness by cognitive scientists, linguists, and other researchers. Some of their concerns are well known: competence theories lack an appreciation of linguistic performance and of the communicative function of language, and they place a strong emphasis on symbolic processing and innateness (see, e.g., [8, 33, 17] for criticisms).

Here we focus on a particular criticism: traditional models fail to acknowledge how much of linguistic structure emerges from communication and embodiment. Recent research on natural language pragmatics, for instance, has focused on language as a cooperative phenomenon where communication is viewed as a *joint action* between the participants [4]. This view is in contrast to the traditional approach in which speaking and hearing are investigated in isolation as *individual actions*. Researchers in the

framework of *emergentism* have argued that the structure of language should be explained as the emergent result of the many interactions between known processes in evolution, development, speaking, listening, and language change over time [17].

This type of work emphasizes the role of (i) the function of language for communication between individuals (*cooperativity*), and (ii) the biophysical constraints of the human body and its environment (*embodiment*) in the explanation for the origin and development of linguistic structure. We are sympathetic to these arguments and share the criticism of a tradition that in some sense equates the *formalisms* of the researcher with the *mechanisms* of the real brain. However, we regret that this general criticism goes hand in hand with a reluctance to use formal models at all. Many researchers have focused instead uniquely on empirical or philosophical approaches (e.g., [17]), or on building "embodied" robots (e.g., [32]).

The goal of this article is to argue that formal models can deal in a meaningful way with embodiment, situatedness, and self-organization. They can help to define these concepts and elucidate the role they play in the development of complex language. *Language games*, such as those studied in recent years in the field of artificial life (see, e.g., [29, 15] for reviews), are a prime candidate for this purpose. Language games are models of language change and language evolution in populations of communicating individuals. Although in most of these models cooperativity and embodiment have not played much of a role, we believe they can be successfully extended to incorporate these important aspects.

The notion of embodiment comes in different flavors. On the one hand, a learning system can be incorporated into an actual robotic body, highlighting the need of the system to cope with sensory limitations [32] and allowing it to manipulate its environment and to develop representations based on sensorimotor interactions with this environment [22]. On the other hand, and more in line with the notion adopted here, embodiment can mean incorporating constraints from sensory, brain, and psychological processing into models without explicitly constructing an artificial body. These approaches are complementary, and neither presents a fully embodied system. In this article we argue that the latter notion of embodiment can be studied with formal models, by incorporating sensory constraints (in the form of noise on the signals) and brain and cognitive processing constraints (by assuming limited processing resources and topological relations between meanings and between signals) into such models.

The models of language evolution that we will consider are *multi-agent models*. They define a population of individuals that talk to each other and learn from each other, using a language that as a result changes over time. Individuals in the models have limited production, memory, and perception abilities, and they have limited access to the knowledge of other individuals. The models evaluate the complex relationship between (i) acoustic, cognitive, and articulatory constraints, (ii) learning and development, (iii) cultural transmission and interaction, (iv) biological evolution, and (v) the complex patterns that are to be explained: the phonology, morphology, syntax, and semantics that are observed in human languages.

The type of language game we examine here is concerned with how a common lexicon can develop in a population of individuals (often called *agents* in this context). In these games, an agent can act either as a speaker or as a hearer. The purpose of a communicative act is the transmission of a meaning from the speaker to the hearer. Meanings cannot be transmitted directly but are encoded by linguistic forms. We can investigate how, based on a great number of such linguistic exchanges under different constraints, a shared lexicon develops so that different speakers use the same word for the same meaning and hearers interpret words with intended meanings. In our models we restrict ourselves to the development of a common lexicon, thus skipping the much more complex and controversial issues in syntax. Nevertheless, we hope

to make the point that language games offer an appealing framework to study other aspects of language as well. For language games that do incorporate grammar, we refer to the extensive review by Kirby [15].

From the perspective of language games, the development of a shared lexicon simply cannot be studied in isolation within one individual, because it depends on the interactions between individuals. In that respect it is a prime example of an aspect of language that escapes study in traditional approaches.

In the rest of this article we will discuss the general framework of these models and present a measure for the quality of a lexicon. We will then study a model that is simple, but is nevertheless novel and serves well to illustrate our approach. Finally, we will discuss how simple language games can be extended to incorporate realistic aspects of cognition, embodiment, and communication.

2 The Optimal Lexicon

The communicative success of a population depends on the organization of the linguistic forms in that population's language, and on how these forms relate to different meanings: how uniquely does one form refer to one meaning? How likely is a speaker to choose a specific linguistic form for a meaning, and how likely is a listener to attribute a certain meaning to a received form? To what extent do individuals agree on the meaning-form mappings? How easily can different forms be confused when communication is noisy?

In this section we will first derive a formal description of what would be the *optimal lexicon*, that is, the lexicon that leads to the highest communicative success in the population. To do so, we need a measure for communicative success. Such a measure is presented next. Similar formalisms were used in [11, 21] and other papers, but our measure is chosen so that we can incorporate some real-world constraints on noise in signaling (like [18]) and different values for different meanings ([14, 19] incorporate in their models the related idea of different frequencies for different meanings).

Speakers can express what they want to say in different ways. Likewise, hearers can interpret spoken forms in different ways. Communicative success is high when the hearer's interpretation of a received form matches with the intention of the speaker. We assume a set of N agents that communicate by forms F to convey meanings M. In a given interaction, a speaker chooses a form f for a meaning m, and the hearer interprets the heard form f^* (which may differ from f if transmission is noisy) and assigns it the meaning m^*. Communication is optimal if speakers and hearers always agree on the meaning for an exchanged form, that is, if $m = m^*$ for any choice of m.

We denote by $S^i(f \mid m)$ the probability that an agent i uses the form f to express the meaning m. Similarly, $R^i(m \mid f)$ is the probability that agent i as a hearer interprets the form f as the meaning m. We assume that there are a finite number $|M|$ of relevant meanings and a finite number $|F|$ of forms used. Further, we assume that similarity between different forms and between different meanings can be measured (e.g., [16]).

We also assume that communication is noisy, that is, the hearer can misperceive a certain form, and more similar forms are more easily confused. We denote by $U(f^* \mid f)$ the probability that an agent perceives the form f as the form f^* (f can be equal to f^*, indicating that the hearer has perceived the form correctly).

Finally, we assume that the communication is successful if the hearer's interpretation is close to the sender's intention. The probability of successfully conveying a certain meaning thus depends on the probabilities of the sender using certain forms and the probabilities of the hearer perceiving and interpreting these forms correctly. We denote by $V(m^*, m)$ the value (or reward) for the hearer understanding m^* when the speaker intended m. Thus V is a measure of communication quality. It should express both

the relative importance of a certain meaning, and the relations between alternative meanings. For example, we could assume that interpreting a signal with a meaning that is wrong but similar is better than interpreting it with just a random meaning, or that being able to express frequent meanings is more important than being able to express infrequent ones.

From these observations, we derive a simple equation that describes the probability $P(m^* \mid m)$ of any hearer j having an interpretation m^* when the speaker i intended m:

$$P(m^* \mid m) = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \sum_{f \in F} \sum_{f^* \in F} \left(S^i \left(f \mid m \right) \cdot U \left(f^* \mid f \right) \cdot R^j \left(m^* \mid f^* \right) \right) \tag{1}$$

This equation says that the probability of the meaning m being perceived as m^* ("understanding m as m^*") is the probability of agent i using the form f to encode meaning m, the hearer perceiving form f^* and then interpreting it as m^*. Because we sum over all N agents as speakers and all but one as hearers ($N-1$; agents do not talk to themselves), we divide the whole expression by $N(N-1)$.

From here it is only a small step to define the communicative success C of the whole population of N agents talking about all $|M|$ meanings:

$$C = \frac{1}{|M|} \sum_{m \in M} \sum_{m^* \in M} \left(P \left(m^* | m \right) \cdot V \left(m^*, m \right) \right) \tag{2}$$

That is, overall communicative success is the sum of the probabilities for all meaning transmissions weighted by their values (assuming that all meanings are equally frequent). This measure is normalized with the number of meanings.

Because S, R, U, and V can all be described as matrices, we can in fact summarize Equations 1 and 2 as follows:

$$C = \frac{1}{|M|N(N-1)} \sum_i \sum_{j \neq i} \left(S^i \times \left(U \times R^j \right) \right) \cdot V \tag{3}$$

where the '$\times$' indicates usual matrix multiplication, and the '$\cdot$' indicates the summation of the product every element in one matrix with its corresponding element in the other matrix (dot multiplication).

Equation 3 constitutes a very general quality measure for a communication system between individuals (described by the matrices S and R), under some *embodied* constraints of articulation and perception (described by U) and semantic/pragmatic constraints on how useful an interpretation is given a certain intention (described by V). By choosing the proper U and V, a wide range of different noise and reward functions can be modeled. However, these matrices can of course not capture all aspects of the embodiment and environment. For instance, the development of conceptual and articulatory abilities and the dependence of rewards and confusion probabilities on specific contexts cannot be modeled directly with our four matrices. However, the formalism is easily extendable to incorporate such aspects. Moreover, even if not all aspects of animal (e.g., [24]), human, or robot communication (e.g., [32]) are modeled, the formalism gives a principled way to abstract out those aspects of embodiment that are nonessential for the emerging language.

With equation 3 in hand, we can now investigate under which conditions communicative success is maximized. We will not provide analytical results for any specific

choice of U and V. Instead, we will present numerical results for a variety of choices of U and V with a simple hill-climbing algorithm. The algorithm used throughout this section is the following:

1. Initialize a population of P individuals, each with an $|M| \times |F|$ matrix S (a production lexicon) and an $|F| \times |M|$ matrix R (a reception lexicon) set with random values and columns normalized.
2. Measure C according to Equation 3.
3. Apply a random change (from a Gaussian distribution with mean 0 and standard deviation $n = 0.1$) to a random entry in a random matrix of a random individual, and normalize the column.
4. Measure C' according to Equation 3.
5. If $C > C'$, revert the change; otherwise $C := C'$.
6. If maximum steps are reached, stop; otherwise go to 3.

Note that in the simulations that use this algorithm an individual's lexicon is not changed as a direct consequence of communication, but is changed randomly. However, this random change may lead to higher communicative success, in which case the change is retained. We use this simple *global optimization procedure* to analyze what the optimal lexicon will look like for different choices of U and V. In Section 3 we will look at the more realistic situation where agents optimize their individual communicative success, that is, where optimization is *local* and *distributed*.

2.1 Categorical Meanings; Noise-Free Signaling

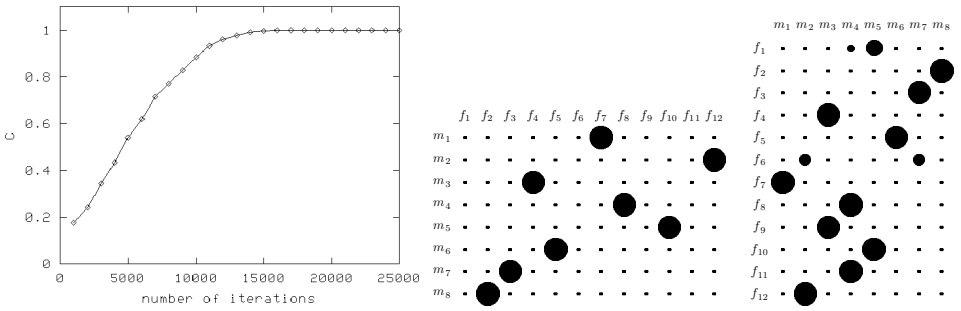
Let us first consider the simplest case of categorical, noise-free communication. That is, we assume that every meaning is unique and has no relation with other meanings. Further we assume that forms are perceived as they are uttered. In short, both U and V are unit matrices (matrices with 1's on the diagonal, and 0's everywhere else).

If we optimize a population's lexicon under these conditions using the hill-climbing algorithm described above, we obtain results as in Figure 1. Here C increases steadily and reaches the optimal value (1.0). The S matrices in the population have maximal probability (= 1.0) for a specific form (horizontal) for each of the meanings (vertical), and probability 0 for all other forms. In the matrix R these forms (vertical) are interpreted as the "correct" meanings. Because there are more possible forms than meanings, some forms are never used and have arbitrary interpretations.

From this simple simulation we can derive two properties of the optimal lexicon: *specificity*, one unique form for every intention, and one unique interpretation for every used form, if $|M| \leq |F|$; and *coherence*, that is, everyone in a population uses the same form for the same meaning.

2.2 Categorical Meanings, Noisy Signaling

If there is noise on the signal (due to a noisy environment and sensory limitations of the hearer), we can expect the hearer to sometimes hear a different form than the speaker uttered. We can model this by introducing nonzero off-diagonal entries in the matrix U. Here, we consider only the simplest case, where forms vary on one axis, determined by their index, and we set the values of U depending on the distance from



(a) Development of communicative success over 25000 iterations

(b) S matrix of a random individual, showing for each meaning (vertical) the probability that she will use any of the forms (horizontal) to

(c) R matrix of the same individual, showing for each form (vertical) the probability that she will choose any of the meanings (horizontal) as its

Figure 1. The optimal lexicon in a population under categorical, noise-free conditions. The size of circles is proportional to the value of the corresponding entry; entries with value 0 are plotted as a small dot. (V and U are unit matrices, $|M| = 8$, $|F| = 12$, $N = 3$, $n = 0.1$).

the "correct" form (and subsequently normalize every row of U):

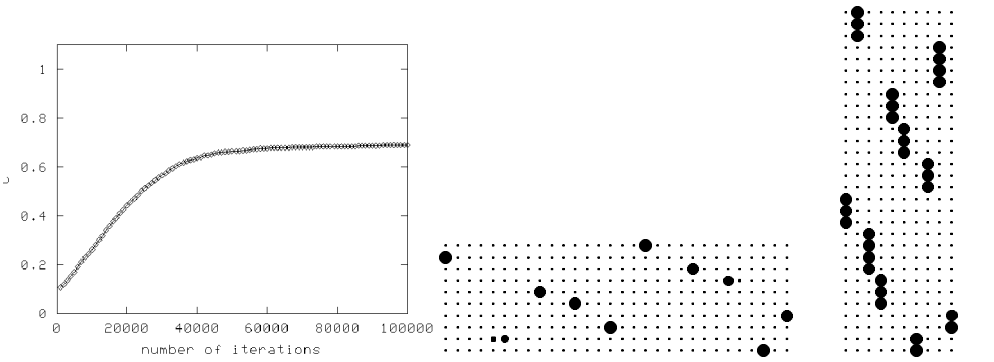
$$U(f^* \mid f) = \frac{1}{1 + (f - f^*)^2} \tag{4}$$

We expect a lower optimal value of C. Moreover, for optimized C, we also expect to find matrices that somehow minimize the chance of misinterpretation. Figure 2 shows that this is indeed what happens. The S matrix shows that for every meaning, there is a prototype form that individuals use. For these prototype forms and their direct neighbors, the interpretation is the "correct" meaning. Thus, little clusters of neighboring forms are all interpreted in the same way, such that prototype forms are maximally distinct from each other. Thus, in addition to specificity and coherence, *distinctiveness* is a property of the optimal lexicon when the signaling is noisy. Note that, even though there are many more forms than meanings, all forms have a specific "best" interpretation. We can obtain similar results with form spaces that have more dimensions [36] or continuous values [37].

2.3 Semantic Similarities and Noisy Signaling

If we include in the model the assumption that not only forms have similarity relations, but also meanings relate to each other, we can identify a fourth criterion of the optimal lexicon: *regularity*. Figure 3 shows results that are obtained by running the hill-climbing algorithm of this section, with U as in Equation 4 and, similarly, V as follows (and rows subsequently normalized):

$$V(m^*, m) = \frac{1}{1 + (m - m^*)^2} \tag{5}$$



(a) Development of communicative success over 100000 iterations

(b) S matrix of a random individual.

(c) R matrix of the same individual.

Figure 2. A local optimum of the lexicon in a population under categorical, noisy conditions (V-unit matrix, U as in Equation 4, $|M| = 10$, $|F| = 30$, $N = 3$, $n = 0.1$).

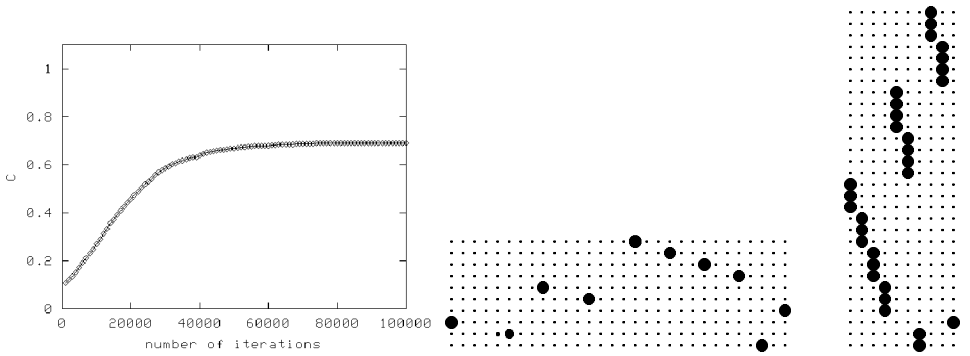
Here, V is maximal when the intended meaning m and understood meaning m^* are the same and decreases with increasing distance between m and m^*.

The local optima found by the hill-climbing algorithm show not only specificity, coherence, and distinctiveness, but also *partial regularity*: similar forms tend to have similar meanings, such that misinterpretations are still better than a random interpretation. The solution found is a local optimum; the globally optimal lexicon is maximally regular: with the parameters of the simulations in Figure 3, meaning m_1 is expressed with form f_1, and forms f_2 to f_3 are interpreted as m_1; meaning m_2 is expressed with f_5, and f_4 to f_6 are interpreted as m_2; and so on. This optimum is not found in this simulation; however, in the local optimum of Figure 3 neighboring clusters of forms are, with only a few exceptions, associated with neighboring meanings. In related work [36] we found that with a slightly different representation the optimum can easily be found as well. Measuring the degree of regularity (as the correlation between the distances between each pair of meanings and the distances between their associated forms) shows that it is consistently higher under conditions with semantic similarities than without.

2.4 Properties of the Optimal Lexicon

From these experiments we can conclude that the optimal lexicon must have the following properties (provided that $|M| \leq |F|$, and that the off-diagonal U and V values are sufficiently low):

- *Specificity*: Every meaning has exactly one form to express it, and vice versa (i.e., there are no homonyms, and no real synonyms: if different forms have the same meaning, they are very similar to each other).
- *Coherence*: All agents agree on which forms to use for which meanings, and vice versa.
- *Distinctiveness*: The forms used are maximally dissimilar to each other, so that they can be easily distinguished.



(a) Development of communicative success over 100000 iterations (b) S matrix of a random individual (c) R matrix of the same individual

Figure 3. Local optima for S and R under semantic similarities, noisy signaling conditions (V as in equation 5, U as in Equation 4, $|M| = 10$, $|F| = 30$, $N = 3$, $n = 0.1$).

- *Regularity*: In the mapping between meanings and forms there is a *preservation of topology*, that is, similar forms tend to have similar meanings.

3 Language Games

After establishing the properties of an optimal lexicon, we can now turn to *language games*, where there is no global optimization, but rather, every individual tries to optimize its own communicative success. Language game models can be viewed as an extension of the basic communication model that consists of a sender, a message, and a receiver. Language games consider a *population* of individuals (*agents*) that can both send and receive. A language game then is a linguistic interaction between two or more agents that follows a specific protocol and has varying degrees of success. The types of models that we will consider have the following components: (i) a linguistic representation, (ii) an interaction protocol, and (iii) a learning algorithm. In this section we will discuss the choices we have made for each of these components, based on a review of existing models.

3.1 Linguistic Representation

By a *representation* we mean here a formalism to represent the linguistic abilities of agents, ranging from recurrent neural networks [1] or rewriting grammars [13, 35] to a simple associative memory [11, 21, 28, 20, 6, 12, 26], representing the strength of associations between meanings and forms.

In the model of this section, we use the same S and R matrices as in Section 2. Forms and meanings thus remain abstract. Other researchers (e.g., [30, 3]) have chosen more concrete representations, such as random concatenations of consonants and vowels for the forms, or positions in a psychophysically motivated color space for meanings. However, these models do not have similarity relations between forms or between meanings. Instead, forms and/or meanings are categorical, and as a result the form-meaning associations in the emerging languages are completely arbitrary (as in our first model, Sections 2.1 and 2.2). A possible exception is the model in [31]; however, in that article it is not clear whether the stochasticity in the meaning space is dependent on the assumed topology (i.e., whether a wrong but close interpretation is more valuable

than a far-off interpretation), and regularity and distinctiveness are not measured or analyzed.

In contrast to these models, we assume here that there are varying degrees of similarity between forms and between meanings, i.e., there is a topological space of meanings, and a topological space of forms. In that respect, our model is more similar to models of the evolution of grammatical language, where associations between structured meanings and structured forms are not arbitrary (e.g., [14, 2]). For the sake of simplicity, we report here results from simulations where forms and meanings each vary on a one-dimensional axis. As in Section 2, we interpret the index of meanings and forms in the S and R matrices as their positions on these axes. Even such a similarity metric, which is only a first step toward more cognitive plausibility, brings fundamentally new behaviors.

3.2 Interaction Protocol

The agents in language game models interact following simple protocols. In most models two agents—a speaker (initiator) and a hearer (imitator)—are chosen at random. Three types of games can be distinguished. In the *imitation game* [5], in contrast to the present models, meanings play no role. However, as in our model and in contrast to most other language game models, the imitation game assumes noise and similarities in the form space and studies the emergent maximization of the distance between them.

In the imitation game, the initiator chooses a random form from its repertoire and utters it. The imitator then chooses the form from its own repertoire that is closest to the received form and utters it. If the initiator finds that the closest match to this (heard) form is the form that it originally used, the game is successful. Otherwise the game is a failure.

In the *naming game* [28], meanings do play a role. The speaker chooses a meaning and a form to express that meaning, and the hearer makes, based on the perceived form, a guess of what is meant. The hearer then receives feedback from the speaker on the intended meaning, that is, whether its guess was correct. The game is a success if the speaker's intention and the hearer's interpretation are the same, and a failure otherwise. The naming game serves as a model system for studying the emergence of conventional form-meaning associations.

In the *observational game*, the meaning of the expressed form is immediately available to the hearer (as in situations where the speaker points at the object that is the topic of a conversation). This simplification has been used in most language game models studied so far (e.g., [11, 28, 21, 1, 13, 12]).

In the model described here, we make another simplifying assumption. We pick two random agents from the population. The first agent learns from the other, and is randomly assigned the role of either speaker or hearer. We then assume that the first agent is able to assess the *overall* communicative success in communicating with the other agent, and learns through a form of hill climbing as described below. The effect of one interaction in our model can thus be seen as the average effect of many interactions in the naming game. In Section 4 we will discuss the consequences of relaxing this assumption.

3.3 Learning Algorithm

In most models, the learning algorithm that agents use to improve their linguistic abilities is very simple (see [27] for a discussion of the required biases of these learning algorithms and how these biases can evolve). In all of the language game models mentioned above, a mechanism is implemented to keep track of the success of each form or form-meaning association. Whether or not a specific association is used depends on this score. Such algorithms can be considered variants of a hill-climbing process:

given a present state of the system, a random variation is tried out. If the performance is better than before, this variation is kept, and otherwise it is discarded.

The difference from a standard hill-climbing algorithm (such as in Section 2) is that optimization is *local* (every agent optimizes its individual success) and many variants are tried out at the same time. That is, at any one time we can view associations with a high score as constituting the present state of the system. For the other associations, the (low) scores are estimates of how much communication would improve by adopting it. If adopting it would improve communication at this point, the scores will go up and the association will eventually become part of the system.

In the language game model of this section, we will simply use a *local* hill-climbing variant. After picking two random agents, the learning agent makes a random change in its S matrix (if it is assigned the role of speaker) or R matrix (if it is the hearer). The learner checks if that change improves the communicative success in communicating with the other agent according to the following equation (which is almost identical to Equation 3, but now for one specific speaker and hearer):

$$C^{ij} = \frac{1}{|M|}\left(S^i \times \left(U \times R^j\right)\right) \cdot V \tag{6}$$

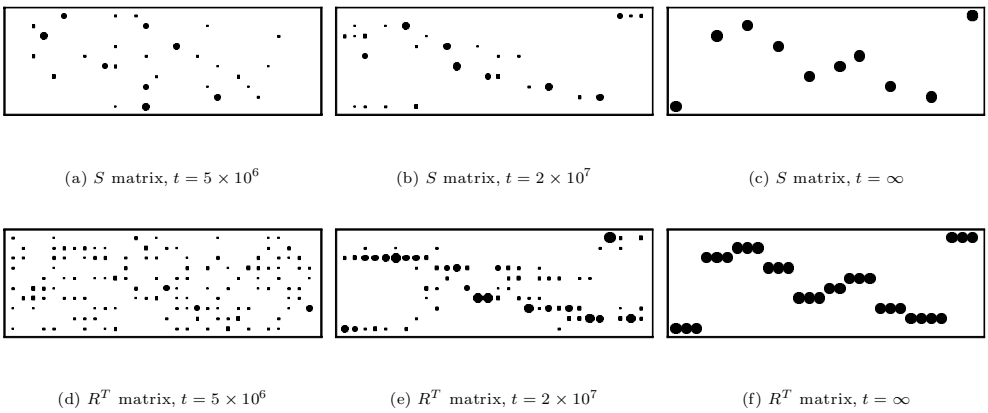
If $C^{ij}_{\text{before}} > C^{ij}_{\text{after}}$, the change is kept; if not, the change is reversed. Note that in this *distributed hill climbing*, at every interaction the target of the hill-climbing process can be different, because each interaction is with a random other agent in the population and because other agents are learning at the same time.

3.4 Self-Organization of the Optimal Lexicon

The main result that we present here is that close approximations of each of the properties of the optimal lexicon emerge from the local interactions that we have defined above. Figure 4 shows results from a simulation with the same parameters as in Figure 3, just with a larger population ($N = 40$) and a higher noise level (the random change in the hill-climbing algorithm is from a Gaussian distribution with mean 0 and standard deviation $n = 1.0$). The figure shows S and R matrices from one random individual at three points in the simulation: after 5×10^6 and 2×10^7 iterations, and in the stable equilibrium configuration (after almost 1×10^8 iterations).

The lexicon that develops shows all four characteristics. In the S matrix at equilibrium (labeled $t = \infty$), every meaning is always expressed by one unique form; in the R matrix, that form is always interpreted with the correct meaning (specificity). At equilibrium, all agents have the same S and R matrices (coherence). In the S matrix, the total distance between all preferred forms is (almost) maximal; in the R matrix, each of these preferred forms (except at the edges) is the center of a little cluster of forms that are all interpreted with the same meaning (distinctiveness). Finally, with three exceptions, all form clusters have neighboring form clusters that express a neighboring meaning (regularity).

The degree of regularity in this simulation is small (the correlation between the distance between each pair of meanings and the distance between their corresponding forms is around 0.2). In general, regularity can be difficult to obtain because to go from an irregular to a regular lexicon many changes to the lexicon are required. Moreover, its contribution to the communicative success is small in comparison with the other three properties. In [36] we show results with a different representation, where the entries in the S and R matrices are always 1 or 0, and random changes move a 1 to a different position in the matrix. In this setup regularity can much more easily emerge, both in the global and in the distributed hill-climbing condition.



(a) S matrix, $t = 5 \times 10^6$ (b) S matrix, $t = 2 \times 10^7$ (c) S matrix, $t = \infty$

(d) R^T matrix, $t = 5 \times 10^6$ (e) R^T matrix, $t = 2 \times 10^7$ (f) R^T matrix, $t = \infty$

Figure 4. Development of specificity, coherence, distinctiveness, and regularity in the lexicon of a population under semantic similarities, noisy signaling conditions. At each time step a random speaker interacts with a random hearer and one of them performs a single hill-climbing step to improve the communication. In this graph, the R matrices are transposed, so that in both S and R^T meanings are on the vertical axis and forms on the horizontal axis. The size of circles is proportional to the value of the corresponding entry; entries with value 0 are not plotted. $t = \infty$ indicates any time after the simulation has converged (from around $t = 10^8$) to the stable equilibrium. (V as in Equation 5, U as in Equation 4, $|M| = 10$, $|F| = 30$, $N = 40$, $n = 1.0$.)

4 Toward More Cognitive Plausibility

Our results show that there is no necessity for explicit and innately specified "principles" that guarantee specificity, distinctiveness, coherence, and regularity. It is possible in principle that these basic characteristics emerge from simple interactions between agents, a generic learning algorithm, and topological meaning and form spaces. That is, they emerge from the embodiment (i.e., general perceptual and processing constraints) and situatedness (i.e., interactions between individuals) of the simulated agents.

Of course, the biophysical constraints of real humans are different from the ones implemented in this model. The next step in our research is therefore to evaluate whether more *realistic* constraints lead—through similar dynamics—to an emergent language with more *realistic* characteristics. Here we consider three possible extensions of the model.

4.1 Limited Feedback

In the distributed hill-climbing simulations we assumed that an agent makes a random change in one of its matrices, and then evaluates if that change increases the success in communicating with one other individual. In reality, that information might not be available. It is therefore worth examining if the same results can be obtained with the minimal assumptions of feedback on whether or not a communication about a single meaning has been successful (as in the naming game [28]), or on shared contexts between speaker and hearer (as in the observational game [25]).

We have done some experiments that show that at least specificity, coherence, and distinctiveness can easily emerge in a naming game setup [37]. Figure 5 shows one of the emerging languages from these experiments. It shows a pattern formed through local interactions between two communicating agents, expressing nine different meanings with forms from a two-dimensional form space. Each of the nine clusters in this figure shows strong associations from two agents for one particular meaning.

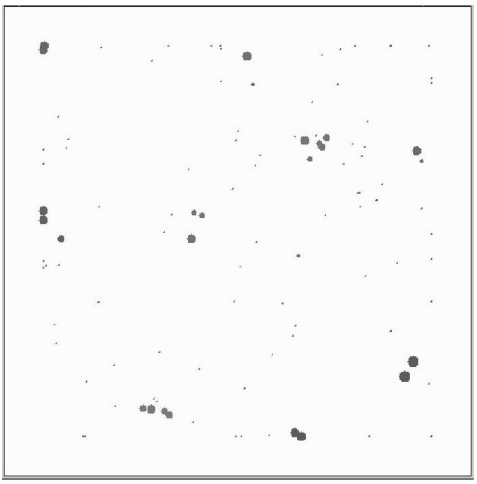


Figure 5. Local interactions: emergence of distinctiveness, coherence, and specificity. Dispersed forms in form space, obtained through local interactions between communicating agents. Each of the nine clusters in this figure shows associations from both agents for one particular meaning. Large dots are strong association. (Parameters: $N = 2$, $|M| = 9$; form space continuous—i.e., $|F| = \infty$; perceptual noise 10%.)

4.2 Cooperativity

An important principle in line with the joint-action view of human communication has been formulated by Grice [9] as the *principle of cooperation*: In a conversation, the speaker makes certain assumptions about the expectations of the hearer, and she uses these assumptions to communicate her intended message effectively. This principle involves the provision of enough, but not too much, information in a message, the relevance of the message to the current conversation topic, and the truthfulness of the information provided. In interpreting the message, the hearer relies on the speaker to have obeyed these principles.

In the context of language game models, we can extend this principle to the cooperative creation of new words: a speaker that is interested in communicative success should only generate a new form if no form for the intended meaning already exists in the language. For example, a speaker who wants to talk about a duck-billed platypus but has forgotten the name for it (or never knew it) would not make up a random word and thus confuse the hearer. Instead, she would either circumvent the term or describe the animal, and somehow prompt the hearer to give the name. By querying the hearer for a possible form, the speaker allows herself to make assumptions about the beliefs of the hearer and therefore to engage in a *cooperative* language game (as opposed to the merely *interactive* language games that are traditionally studied). Such an extension of the language game framework is plausible in that it views language as a cooperative phenomenon and as a means to maximize the efficiency of communicating intended meanings. It will prevent the creation of an excess of new forms, thereby reducing the number of synonyms and the cognitive load.

4.3 Analogy

When an agent creates a new form in a language game, it usually randomly assembles phonemes (e.g., [28]). This mechanism is in line with the claim of the "arbitrariness of the sign" [7]: the structure of the form has no relationship to the meaning conveyed by it. While this is true for many forms in today's existing languages, there is evidence suggesting that, in the creation of new forms, the intended meaning should be taken

into account. First of all, when new words are created in, for example, English, they are often compounded and derived from existing words to ease their understanding. Thus, someone who eats bananas will be called a "banana-eater" rather than a "manslo," to indicate the semantic relationship with bananas and eaters. While such a process cannot be applied to simple language games directly, it does show a structural relationship between words that reflects a semantic relationship between their meanings.

Second, there is growing evidence for the hypothesis that the sound of a word can suggest its meaning ("sound symbolism"). This idea was first mentioned by Plato and has been pursued since then, for example, by von Humboldt [34]. Subsequent psycholinguistic research has shown that in the formation of words, certain sounds can represent certain meanings. For example, in assigning the two words *Mil* and *Mal* to images of big and small tables, 80% of subjects chose *Mal* to stand for the larger table and *Mil* for the smaller table, indicating that /a/ suggests large size and /i/ small size [23]. These results have been reproduced and extended by numerous researchers (see e.g. [10]).

A less controversial version than such *absolute* sound symbolism (where sounds carry meaning) is a *relative* sound symbolism that can be directly applied to the creation of new forms in naming games. It is described by von Humboldt [34, p. 74] as "Words whose meanings lie close to one another, are likewise accorded similar sounds," while the sounds themselves bear no direct semantic content. In Sections 2 and 3 we presented results where such relative sound symbolism (regularity) emerges as an optimal solution in noisy conditions. However, we can also imagine that agents actively exploit a form of topology preservation when creating new forms. In a language game the decoding of the form by the hearer could then work as follows:

```
Find a meaning for the form f:
for the nearest neighbor f' of f according to the similarity
    metric, find the best meaning m'
associate f with a meaning which is closest to m'
```

This approach can help to reduce ambiguity in the hearer's lexicon. Preliminary results suggest faster convergence of the language than in the original model, due to the emergence of regularities in the form-meaning mapping. Further, we found several examples of parameter settings that would not lead to convergence under the classical settings, but did converge under these topological settings. Finally, we find an unexpected delay in the convergence in the final stage, due to conflicts between competing partial regularities. This delay indicates that lexicon creation is subject to the opposing pressures of *topological preservation* and *distinctiveness maximization*. We could assume that in the evolution of vocabularies of human languages, words with similar meanings might have developed to be as similar as possible (and thus predictive of their meaning) while at the same time being as distinctive as possible (to facilitate communication with already known words). A new form that is created to be similar to another in order to facilitate understanding of its meaning would then undergo variation (historical change) to become more arbitrary as it became more established and a prediction of its meaning became less important than its distinctiveness from other forms. While we have not incorporated these constraints in our current simulations, we believe that they present a promising direction in the endeavor to integrate language game formalisms with cognitive approaches to language.

5 Conclusions

We have discussed the relevance of language evolution models to the study of embodiment and self-organization of language, and presented a formalism for describing

language games. Language game models are complementary to work that studies language processing and language acquisition. The models we discussed are simple; their value is that they make the roles of diachrony, embodiment, and self-organization in emerging linguistic structure explicit and testable.

We have argued that the environment and embodiment of communicating agents in the real world impose a topology on both the meaning and the form space of their communication system. We have shown that with these topologies the optimal lexicon has four characteristics: specificity, coherence, distinctiveness, and regularity. We have further shown that in a distributed population of agents that each have generic learning capabilities, a lexicon can be established that shows each of these four characteristics.

Our results on distinctiveness and regularity follow naturally from the framework that we have described in this article. Nevertheless, they have not been reported in the extensive literature on the modeling of language evolution. We believe that this fact in itself is support for our approach to embodiment, where we try to incorporate constraints from sensory, brain, and psychological processing into formal models without explicitly constructing an artificial body. However, much work remains to be done on explaining the role of these constraints in the evolution of language. In the final part of the article, we have therefore raised issues where cognitive science can inform language game modeling, and eventually lead to a detailed understanding of how complex language has emerged from many simple interactions.

Acknowledgments

The writing of this article was supported by European Commission RTN grant HPRN-CT-2000-00065 to Gert Westermann, and a Prins Bernhard Cultuurfondsbeurs and a Marie Curie fellowship of the European Commission to Willem Zuidema. Part of the research described in this article was performed while WZ was at the A.I. Lab of the Vrije Universiteit Brussel and funded through a Concerted Research Action fund (G.O.A.) of the Flemish Government and the VUB.

We thank Kenny Smith, Charlotte Hemelrijk, Hanspeter Kunz, and two anonymous reviewers for helpful comments.

References

1. Batali, J. (1998). Computational simulations of the emergence of grammar. In J. Hurford & M. Studdert-Kennedy (Eds.), *Approaches to the evolution of language: Social and cognitive bases*. Cambridge, UK: Cambridge University Press.
2. Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models*. Cambridge, UK: Cambridge University Press.
3. Belpaeme, T. (2001). Simulating the formation of color categories. In B. Nebel (Ed.), *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'01)* (pp. 393–398). San Francisco: Morgan Kaufmann.
4. Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.
5. De Boer, B. (1999). *Self-organisation in vowel systems*. Ph.D. thesis, Vrije Universiteit Brussel AI lab.
6. de Boer, B., & Vogt, P. (1999). Emergence of speech sounds in changing populations. In D. Floreano, J.-D. Nicoud, & F. Mondada (Eds.), *Advances in Artifical Life* (pp. 664–673). Berlin: Springer-Verlag.
7. de Saussure, F. (1916). *Course in general linguistics*. La Salle, IL: Open Court. Translated by Roy Harris. Edition published in 1986.

8. Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisin, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

9. Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, volume 3: Speech acts* (pp. 41–58). New York: Academic Press.

10. Hinton, L., Nichols, J., & Ohala, J. J. (Eds.) (1995). *Sound symbolism*. Cambridge, UK: Cambridge University Press.

11. Hurford, J. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, *77*, 187–222.

12. Kaplan, F. (2000). *L'émergence d'un lexique dans une population d'agents autonome*. Ph.D. thesis, Université Paris 6, Sony CSL-Paris.

13. Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, J. Hurford, & M. Studdert-Kennedy (Eds.), *The evolutionary emergence of language: Social function and the origins of linguistic form*. Cambridge, UK: Cambridge University Press.

14. Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*, 102–110.

15. Kirby, S. (2002). Natural language from Artificial Life. *Artificial Life*, *8*, 185–215.

16. Landauer, T., Foltz, P., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

17. MacWhinney, B. (Ed.) (1999). *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum Associates.

18. Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences of the U.S.A.*, *96*, 8028–8033.

19. Nowak, M. A., Plotkin, J. B., & Jansen, V. A. (2000). The evolution of syntactic communication. *Nature*, *404*, 495–498.

20. Oliphant, M. (1999). The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior*, *7*.

21. Oliphant, M. & Batali, J. (1996). Learning and the emergence of coordinated communication. *Center for Research on Language Newsletter*, *11*.

22. Pfeifer, R. & Scheier, C. (1999). *Understanding intelligence*. Cambridge, MA: MIT Press.

23. Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, *12*, 225–239.

24. Seyfarth, R. M., & Cheney, D. L. (1997). Some general features of vocal development in nonhuman primates. In C. T. Snowdon & M. Hausberger (Eds.), *Social influences on vocal development* (pp. 249–273). Cambridge, UK: Cambridge University Press.

25. Smith, A. D. (2001). Establishing communication systems without explicit meaning transmission. In J. Kelemen & P. Sosík (Eds.), *Advances in Artificial Life (Proceedings 6th European Conference on Artificial Life, Prague)*. Berlin: Springer.

26. Smith, K. (2002). The cultural evolution of communication in a population of neural networks. *Connection Science*, *14*, 65–84.

27. Smith, K. (2003). Natural selection and cultural selection in the evolution of communication. *Adaptive Behavior*, *10*(1), 25–44.

28. Steels, L. (1997). Self-organising vocabularies. In C. Langton & K. Shimohara (Eds.), *Proceedings of the 5th International Workshop on Artificial Life: Synthesis and simulation of living system* (pp. 179–184). Cambridge, MA: MIT Press.

29. Steels, L. (1997). The synthetic modeling of language origins. *Evolution of Communication*, *1*, 1–35.

30. Steels, L. (1998). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence*, *103*, 133–156.

31. Steels, L., & Kaplan, F. (1998). Stochasticity as a source of innovation in language games. In C. Adami, R. Belew, H. Kitano, & C. Taylor (Eds.), *Proceedings of Artificial Life VI* (pp. 368–376). Cambridge, MA: MIT Press.

32. Steels, L., Kaplan, F., McIntyre, A., & Van Looveren, J. (2002). Crucial factors in the origins of word-meaning. In A. Wray (Ed.), *The Transition to Language*. Oxford, UK: Oxford University Press.

33. Tomasello, M. (Ed.) (1998). *The new psychology of language: Cognitive and functional approaches to language structure*. Mahwah, NJ: Lawrence Erlbaum Associates.

34. von Humboldt, W. (1836). *On Language*. Cambridge, UK: Cambridge University Press. Translated from the German by Peter Heath. Edition published in 1988.

35. Zuidema, W. (2003). 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)* (pp. 51–58). Cambridge, MA: MIT Press.

36. Zuidema, W. (2003). Optimal communication in a noisy and heterogeneous environment. 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)* (pp. 553–563). Berlin: Springer Verlag.

37. Zuidema, W. & Westermann, G. (2001). Towards formal models of embodiment and self-organization of language. In *Proceedings of the Workshop on Developmental Embodied Cognition*. Edinburgh, UK.