Semiotic Dynamics Solves the Symbol Grounding Problem

Luc L. Steels[1,2], Martin Loetzsch[1] & Michael S. Spranger[1]

[1] *Sony Computer Science Laboratory - Paris*

[2] *Artificial Intelligence Laboratory, Vrije Universiteit Brussel*

Language requires the capacity to link symbols (words, sentences) through the intermediary of internal representations to the physical world, a process known as symbol grounding[1]. One of the biggest debates in the cognitive sciences concerns the question how human brains are able to do this. Do we need a material explanation or a system explanation? John Searle's well known Chinese Room thought experiment[2], which continues to generate a vast polemic literature of arguments and counter-arguments, has argued that autonomously establishing internal representations of the world (called 'intentionality' in philosophical parlance) is based on special properties of human neural tissue and that consequently an artifical system, such as an autonomous physical robot, can never achieve this. Here we study the Grounded Naming Game as a particular example of symbolic interaction and investigate a dynamical system that autonomously builds up and uses the semiotic networks necessary for performance in the game. We demonstrate in real experiments with physical robots that such a dynamical system indeed leads to a successful emergent communication system and hence that symbol grounding and intentionality can be explained in terms of a particular kind of system dynamics[3–5]. The human brain has obviously the right mechanisms to participate in

this kind of dynamics but the same dynamics can also be embodied in other types of physical systems.

Let us first operationalise under what conditions an agent (biological or artificial) can be said to be capable of symbol grounding. The well known Turing test for intelligence[6] relies on subjective judgement and has been widely criticised, precisely because it does not examine the capability to deal with symbols about the world, which is the main issue addressed here. Picking a concrete task like planning or navigation is also controversial because there is often debate whether it involves the manipulation of symbols or even of internal representations of the world [7]. However, nobody doubts that language is symbolic and that it expresses representations of the world. So we propose to use performance in language games as test criterion. Specifically we focus here on the Grounded Naming Game in which participants draw attention to physical objects in their shared context using names. The Grounded Naming Game is played between two randomly chosen members of a group. All players are considered equal and have an equal chance to interact with each other. The environment contains a set of ten physical objects which differ enough to be uniquely identifiable, but the players do not know in advance how many objects there are nor their characteristic features. They have never seen the objects and so there can be no prior vocabulary of names. A game starts when two players have been able to establish a joint attention frame. They have physically encountered each other, have established their mutual position, and have located some objects in the immediate context. One of them then randomly takes on the role of speaker and the other is the hearer. The speaker chooses one object as topic and names it. The hearer looks up the meaning of this name in his own vocabulary and checks through his own perception of the world which object could have been intended. If he has an opinion, the hearer points to this object and the game is a success if the speaker agrees that this was the topic he had chosen. Otherwise the speaker points to the topic and the hearer gets an opportunity to learn. Of course in the beginning most games will fail, but a group of players capable of intentionality and symbol grounding should gradually be able to settle on a shared vocabulary and reach a very high level of success. It could be argued that this would only prove that players can give intrinsic semantics to *external* symbols, but obviously if they have this capacity they can also use such symbols internally, for example to perform symbolic inference[8]. Some psychological experiments have recently examined how well humans can self-organise symbol systems from scratch in similar game-like interactions[9,10]. They clearly can. But here we are interested to understand what information structures and processing mechanisms are needed *in principle* and whether they depend on the unique causal force of certain materials or are a system property. If the latter is the case, we should be able to program artificial agents to play the game successfully without human intervention, neither by engineering the robots' internal representations or use of symbols directly, nor by teaching an existing human symbol system through supervised learning, because in both cases the semantics is parasitic on human intelligence. This is what we have done using the QRIO humanoid robot[11] as experimental platform (figure 1).

Before examining the dynamics, it is useful to emphasise how extraordinary difficult the task is. First of all it is extremely challenging to set up joint-attention frames, which may explain why no other animals except humans can self-organise shared symbolic systems[12]. Second, it is non-trivial to identify physical objects based on visual sensations, particularly if both the objects and the robots move around. Once object regions are found (figure 2), feature detectors can compute values for color channels, brightness, speed of movement, etc., but these will always be very noisy. Moreover because the robots look at the scene from different points of view, they necessarily have different sensory experiences of the same objects and localise them differently within their own egocentric reference frames. Third, it is very difficult to establish which individual is associated with a particular sensory experience because the appearance of an object changes depending on its posture, its position with respect to the viewer, and the changing light conditions in the environment. The experiment uses a standard method in object recognition, for which there is also evidence in human subjects, which is to capture the invariant properties of a particular view of an object in terms of prototypes[13]. Prototypes can be represented by the mean, variance, and weight of values on each sensory dimension and the best matching prototype can be found by nearest neighbor computation. Many neural network models perform this kind of computation, such as Radial Basis Function Networks[14], but the extra difficulty here is that robots do not have access to a clear data set of examples and counter-examples. Fourth, although an individual object may have invariant properties for one of its views, usually there are significant differences between different views (for example a front view of an object standing up and a back view of the same object laying down) and so how can a robot learn that two views belong to the same object? Finally, the robots must build up a vocabulary associating names with individuals. When a robot does not know a name, he can baptise the object with a newly invented name that spreads in consecutive games if hearers can acquire its meaning. But since a language game is always a local interaction between only two players, another robot may invent a different name for the same object which could also propagate to some extent, and so synonymy (different names for the same object) is unavoidable. Moreover because of the inherent noise and unreliability of real-world vision, it is possible that one robot misinterprets feedback or makes an incorrect guess about the meaning of an unknown name and thus acquires a different meaning. So homonymy (different objects for the same name) is unavoidable as well. How then, can a shared vocabulary be reached without central control or telepathy?

We argue that the solution to these various issues lies in setting up a particular ’semiotic’ dynamics that gradually coordinates sensations, sensory experiences, prototypical views, individuals, and names, both within a single agent and across the population. The framework of complex networks[15], which is playing such an important role in many sciences today is useful to formulate and understand this dynamics. Each agent in the population, $a \in P$, should maintain a semiotic network $\mathcal{S}_a = O_a \times V_a \times I_a \times N_a$ where O_a is the set of sensory experiences of the agent grouped per scene, V_a the set of prototypical views maintained by a, I_a the set of individuals known to a, and N_a the set of names in a’s vocabulary (see figure 3). Each link in the network is weighted (with a real number between 0.0 and 1.0). The weight of the link between a sensory experience and a prototypical view is based on nearest neighbor computation. The other weights are stored in memory and reflect the confidence of the agent in the use of that link based on past experience. A semiotic network is bidirectional and dynamic in the sense that new nodes can be added or removed as a side effect of a language game and the weights between nodes change based on the outcome of a game. In order to decide which name to use for a chosen topic, the speaker traces pathways in his private semiotic network. Starting from the sensory experiences of the objects perceived in the current scene, he activates the best matching prototypical views, activates the individuals linked to these prototypical views, and then looks up the names for them. The pathway that has the highest cumulative score, which is the sum of all weights of the links involved, is the winner and the name occurring at the endpoint of this path is the name transmitted by the speaker to the hearer. Conversely, in order to decide which physical object to point at, given a name, the hearer traces pathways in his own private network but now in the other direction, starting from the name. The object occurring at the endpoint of the path with the highest cummulative score is the topic to which the hearer points. The speaker interprets the pointing gesture and then gives appropriate feedback.

The key question is obviously: What are the operators that are building and changing the semiotic networks in each agent? We focus first on the vocabulary, i.e. the links between individual objects and their names. Dynamical systems for the self-organisation of vocabularies have already been studied extensively using the (non-grounded) Naming Game[16, 17]. Several viable solutions are known. The one used here relies on lateral inhibition, familiar from several bi-directional neural network models[18]. After a successful game, both speaker and hearer increase the weight of the lexical associations involved in the winning path and decrease that of associations with the same individual but a different name, so that there is a damping of synonymy, and with the same name but a different individual, so that there is a damping of homonymy. After an unsuccessful game, only the weight of the chosen association is decreased. When these rules are used collectively in consecutive games between randomly chosen members of the population a vocabulary quickly self-organises due to the positive feedback in the system (Figure 4). This process is reminiscent of many selectionist processes and similar to phenomena studied in opinion and social dynamics[19]. Mathematical proofs of convergence[20], scaling laws[21] and the non-trivial impact of social network structure on naming dynamics[22] have now been well studied, but these investigations assume that all agents know what kind of objects there are in their world *a priori* and that there is perfect shared knowledge of which objects appear in the context of a specific game. The main achievement of the Grounded Naming Game experiment is to take away this scaffold. In order to do so, we need a mechanism explaining where prototypical views and their links to individual objects come from.

When an agent sees a scene in which there are different segments, each yielding their own sensory experience, he can safely assume that these segments belong to different individuals and therefore must match best with different prototypes. If this condition is violated, the agent can use the sensory experiences without a unique match as seeds for new prototypical views and link them to newly introduced individuals. Prototypes are later adjusted to better reflect invariant properties by updating their mean value and variance. Figure 5 shows what happens when this strategy is adopted. The population quickly reaches a high level of communicative success (above 90 %). However the average number of individuals in the agents' semiotic networks is much larger than the ten distinctive objects introduced in the experiment. Apparently, agents are naming prototypical views of individual objects instead of the individuals themselves. We have not adequately addressed how a robot learns to interrelate different views. There is in fact no guaranteed way to learn this, even for humans. However there are various heuristics that can be employed. We have operationalised one concrete example. When an object is moving or being moved, its appearance may change but the observer who is tracking the object, still knows that he is dealing with the same object and so he can exploit this information to fine-tune his semiotic networks. We have endowed the QRIO robots with image processing algorithms for tracking. They align object regions across different images, enriched with top-down predictions of how an object region will change over time based on Recursive Bayesian estimation using Kalman filters[23]. The object being moved by the experimenter in Figure 2 yields two quite different sensory experiences when standing up or lying down which match with two different prototypes, but thanks to the tracking heuristic, the semiotic network can be rearranged to reflect that they are two views of the same object (Figure 6). Figure 7 shows what happens when the robots use this heuristic. The population exhibits still the same capacity to achieve communicative success as in Figure 5, but now the number of individuals has significantly lowered, approaching the experimental target of ten physical objects. Clearly humans use many additional heuristics. For example, if we see somebody walking into a building with a refrigerator and we later see the same person on the top floor handling a refrigerator we will assume it is the same refrigerator, even if we could not track this object. The point here is not to operationalise all imaginable heuristics but to show that heuristics help to optimize and coordinate the semiotic networks between individuals and thus further increase their ability to develop internal representations anchored in the world.

The Grounded Naming Game experiment demonstrates that symbol grounding and establishing intrinsic semantics can be achieved through the coordination of semiotic networks in situated embodied interactions and hence need not depend on special properties of matter as uniquely possessed by the neural tissue of the brain, such as unknown quantum gravity effects of microtubules[24]. We use here QRIO robots but it is of course possible to embody the same principles in other robots, run the software on other processors and other operating systems. Completely different physical objects can be used, perceived through other sensory modalities (like sound or touch) and identified through other object recognition techniques. We can use other update rules for self-organising the vocabulary and additional heuristics for tracking object identity. What is crucial is the overall system, not this particular embodiment. The present experiment constitutes a clear breakthrough in artificial intelligence research because it shows for the first time how robots can self-organise a grounded symbol system. Humans have to deal with the same issues and so this kind of study of semiotic dynamics helps us understand what our embodied minds need to be able to do in order to bootstrap symbolic communication systems grounded in the world.

1. Harnad, S. The symbol grounding problem. *Physica D* **42**, 335–346 (1990).

2. Searle, J. Minds, brains, and programs. *Behavioral and Brain Sciences* **3**, 417–424 (1980).

3. Cangelosi, A., Greco, A. & Harnad, S. From robotic toil to symbolic theft: Grounding transfer from entry-level to higher level categories. *Connection Science* **12**, 143–162 (2000).

4. Steels, L. Intelligence with representation. *Phil Trans Royal Soc Lond A.* **361**, 2381–2395 (2003).

5. Hoffstadter, D. *I am a Strange Loop.* (Basic Books, New York, 2006).

6. Turing, A. Computing machinery and intelligence. *Mind* 433–460 (LIX).

7. Brooks, R. Intelligence without representation. *Artificial Intelligence* **47**, 139–159 (1991).

8. Newell, A. & Simon, H. A. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM* **19**, 113–126 (1976).

9. Galantucci, B. An experimental study of the emergence of human communication systems. *Cognitive Science* **29**, 737–767 (2005).

10. Selten, R. & Warglien, M. The emergence of simple languages in an experimental coordination game. *PNAS* **104**, 7361–7366 (2007).

11. Fujita, M., Kuroki, Y., Ishida, T. & Doi, T. T. Autonomous behavior control architecture of entertainment humanoid robot sdr-4x. In *Proceedings IROS 2003*, 960–967 (2003).

12. Tomasello, M. Joint attention as social cognition. In Moore, C. & Dunham, P. J. (eds.) *Joint Attention: Its Origins and Role in Development* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1995).

13. Edelman, S. Representation, similarity and the chorus of prototypes. *Minds and Machines* **5**, 45–68 (1995).

14. Bischop, C. Neural networks and their applications. *Rev. Sci. Instr.* **65**, 1830–1832 (1994).

15. Strogatz, S. Exploring complex networks. *Nature* **410**, 268–276 (2001).

16. Steels, L. A self-organizing spatial vocabulary. *Artificial Life* **2**, 319–332 (1995).

17. Barr, D. Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science* **28**, 937–962 (2004).

18. Kosko, B. Bidirectional associative memories. *IEEE Transactions on Systems, Man and Cybernetics* **18**, 49–60 (1988).

19. Axelrod, R. & Hamilton, W. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).

20. De Vylder, B. & Tuyls, K. How to reach linguistic consensus: A proof of convergence for the naming game. *Journal of Theoretical Biology* **242**, 818–831 (2006).

21. Baronchelli, A., Felici, M., Caglioti, E., Loreto, V. & Steels, L. Sharp transition towards shared lexicon in multi-agent systems. *J.Stat.Mech* **P06014** (2006).

22. Dall'Asta, L., Baronchelli, A., Barrat, A. & Loreto, V. Non-equilibrium dynamics of language games on complex networks. *Phys.Rev. E* **74** (2006).

23. Kalman, R. E. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering* **82**, 35–45 (1960).

24. Hameroff, S. & Penrose, R. Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness? In Hameroff, S., Kaszniak, A. & Scott, A. (eds.) *Toward a Science of Consciousness - The First Tucson Discussions and Debates*, 507–540 (The MIT Press, 1996).

Acknowledgement The authors are grateful to M. Fujita and H. Shimomura of the Intelligent Systems Research Labs at Sony Corp, Tokyo. They have developed with their team the extraordinary QRIO robots used in these experiments and gave access and support to use these robots for the experiments. This research was carried out at the Sony Computer Science Laboratory in Paris and Tokyo with additional support from the EU-FET ECAgents project (IST-2003 1940).

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to L. Steels (email: steels@arti.vub.ac.be).

Figure 1 Experimental Setup. The QRIO humanoid robot is about 60 cm high and weighs 7.3 kg. Its sensors include two cameras in the head, a microphone, and sensors in each motor joint to monitor posture and movement. The robot has enough computing power and battery to autonomously walk around on its two legs and perform various actions in the world. The software state of an agent is downloaded in a robot's memory and uploaded after each game so that a large population of agents can use the same bodies to interact in the world. The environment consists of uniquely identifiable objects, typically colorful geometric shapes or toys. The experiments have been repeated for different collections of ten physical objects each.

Figure 2 Steps of object recognition and tracking for three points in time. The first column contains the source images. Robots scan the image and classify picture elements (pixels) according to whether they are foreground or background and whether motion has occurred (second column). All regions standing out against the background are considered to be candidate objects. The third column shows the changing histogram of the green-red channel for object o_{716}. This histogram is used to track o_{716} in space and time using Recursive Bayesian estimation techniques (applied in column 4). Knowing the offset and orientation of the camera relative to the body, the robots are able to estimate the position and size of objects in the world (egocentric reference system, applied in column 5). The size of each circle shows the perceived width of the object. Together with the position and orientation of the other robot (black arrows), this becomes the set of sensory experiences of one robot for one scene.

Figure 3 Snapshot of the semiotic network of a single agent. Weighted links connect sensory experiences to prototypical views, views to individuals, and individuals to names. Production and interpretation traces pathways in these networks and chooses a winner-take-all based on the highest cumulative score.

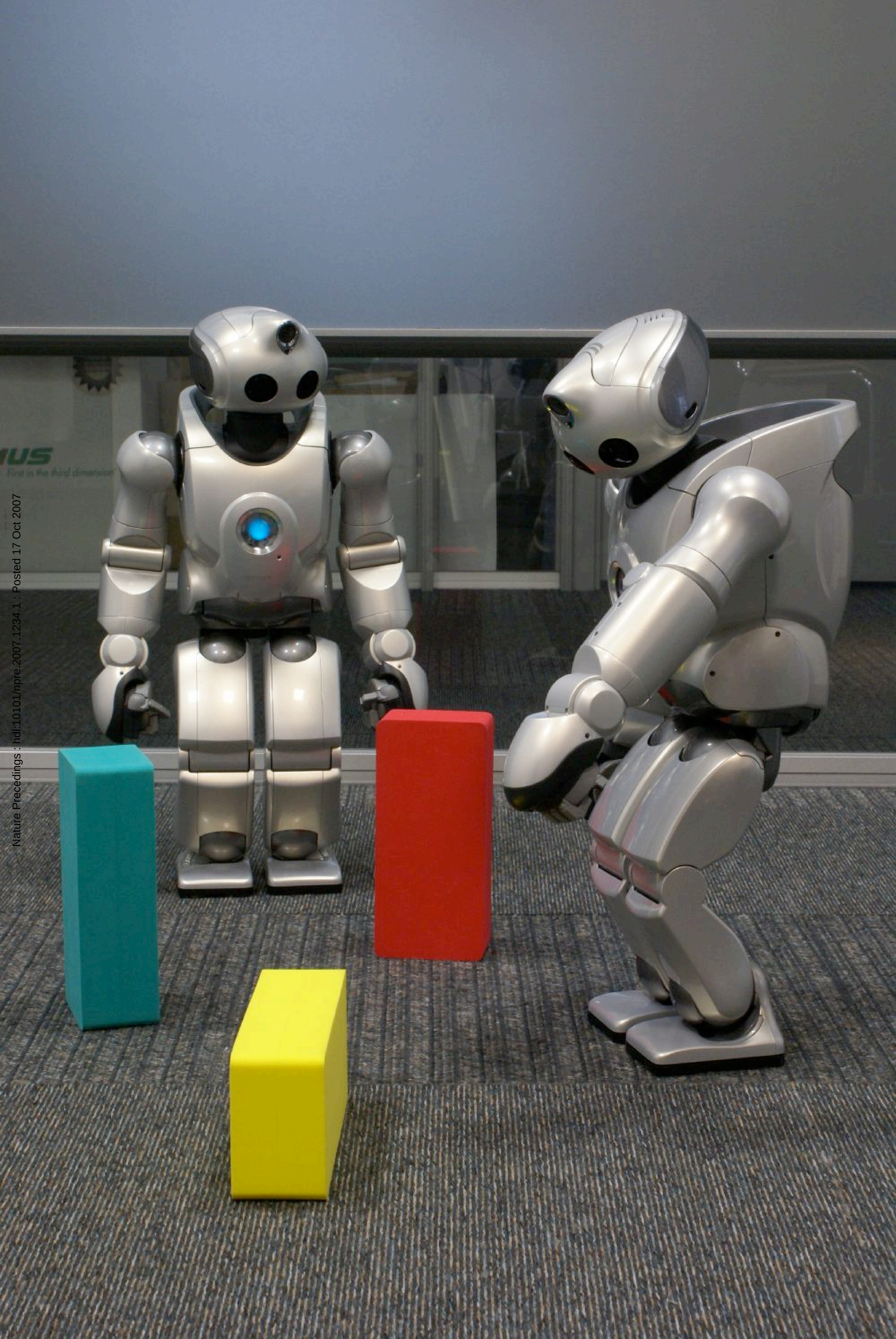
Figure 4 The (ungrounded) Naming Game in a population of 10 agents which have shared a priori knowledge about individuals. The communicative success and average number of names for the complete population is shown on the y-axis. The number of games is plotted on the x-axis. Communicative success rises rapidly to reach total success. The average vocabulary size grows as new words are invented and propagated until all agents know at least one word for each individual object, and then a phase where the vocabulary gradually becomes optimal due to the positive feedback effect.

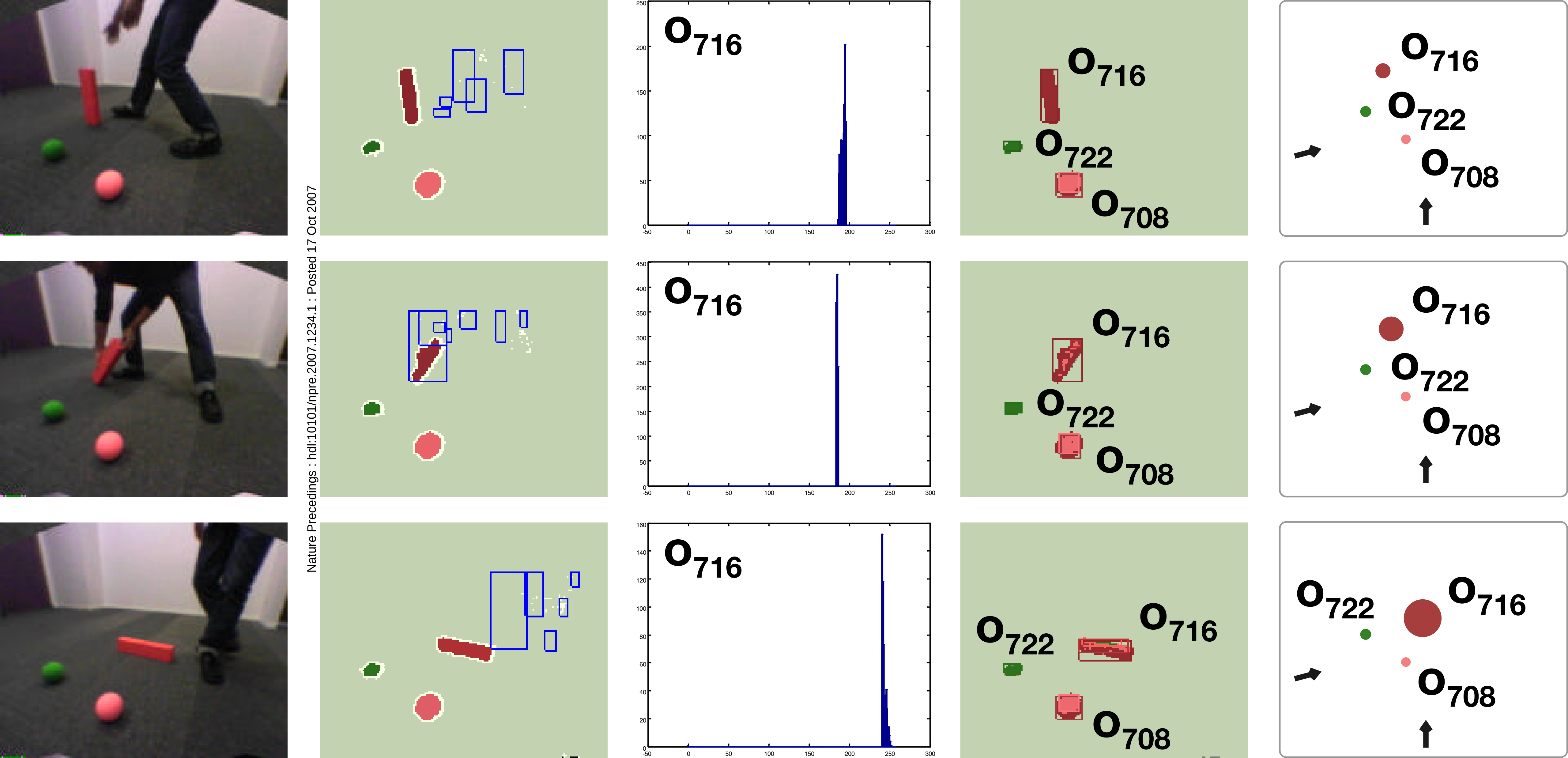
Figure 5 The Grounded Naming Game in a population of 10 agents for a set of 10 unknown individuals. 20,000 language games have been performed, averaged over ten experimental runs. Communicative success, vocabulary size, the average number of individuals and the average number of prototypes for the agents are shown. There are as many prototypes as individuals indicating that agents name prototypes rather than individuals.

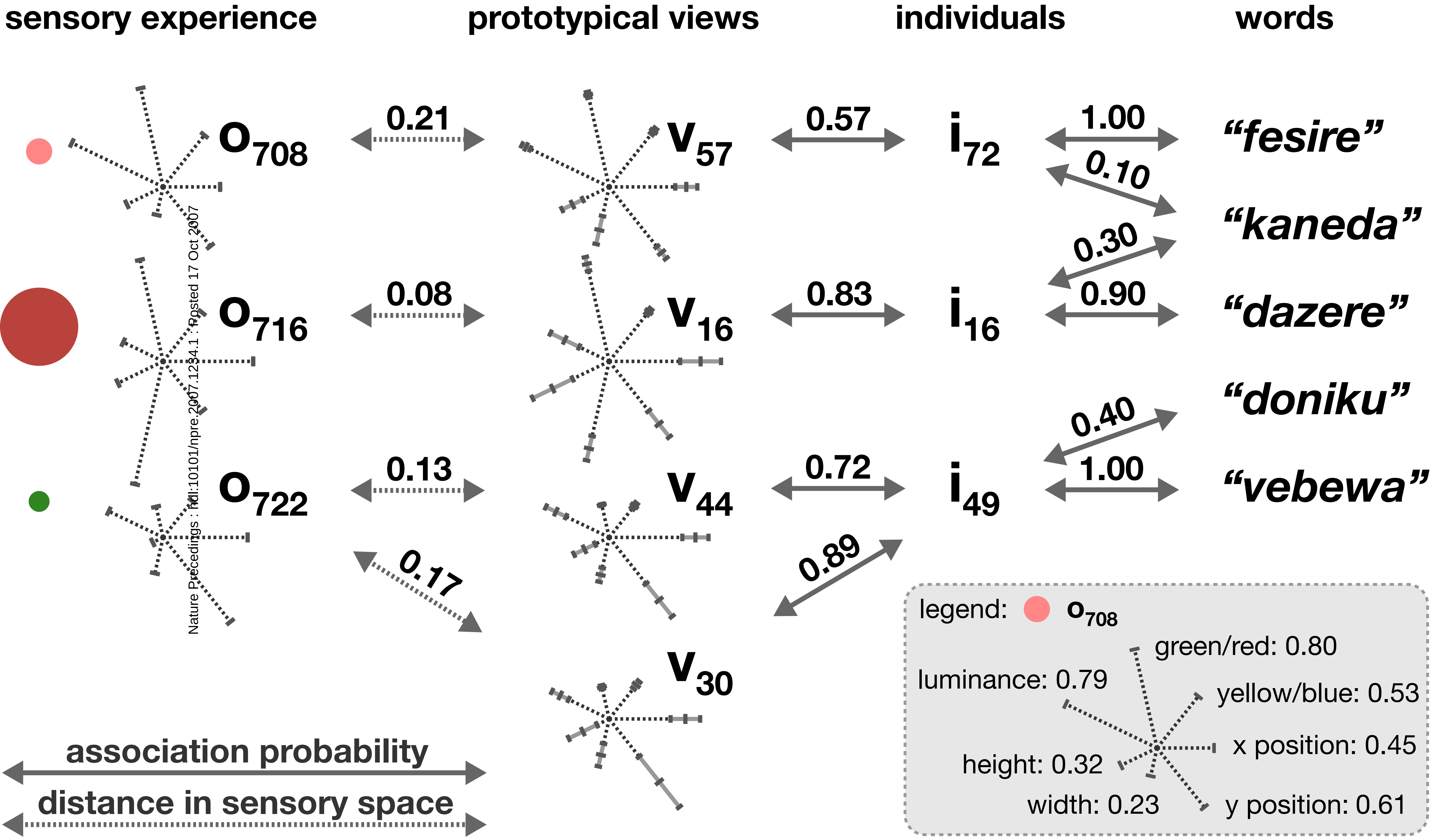
Figure 6 Agents utilise information of object identity obtained through the tracking heuristic in order to reorganise their semiotic networks. The prototypical views v_{73} and v_{16} are

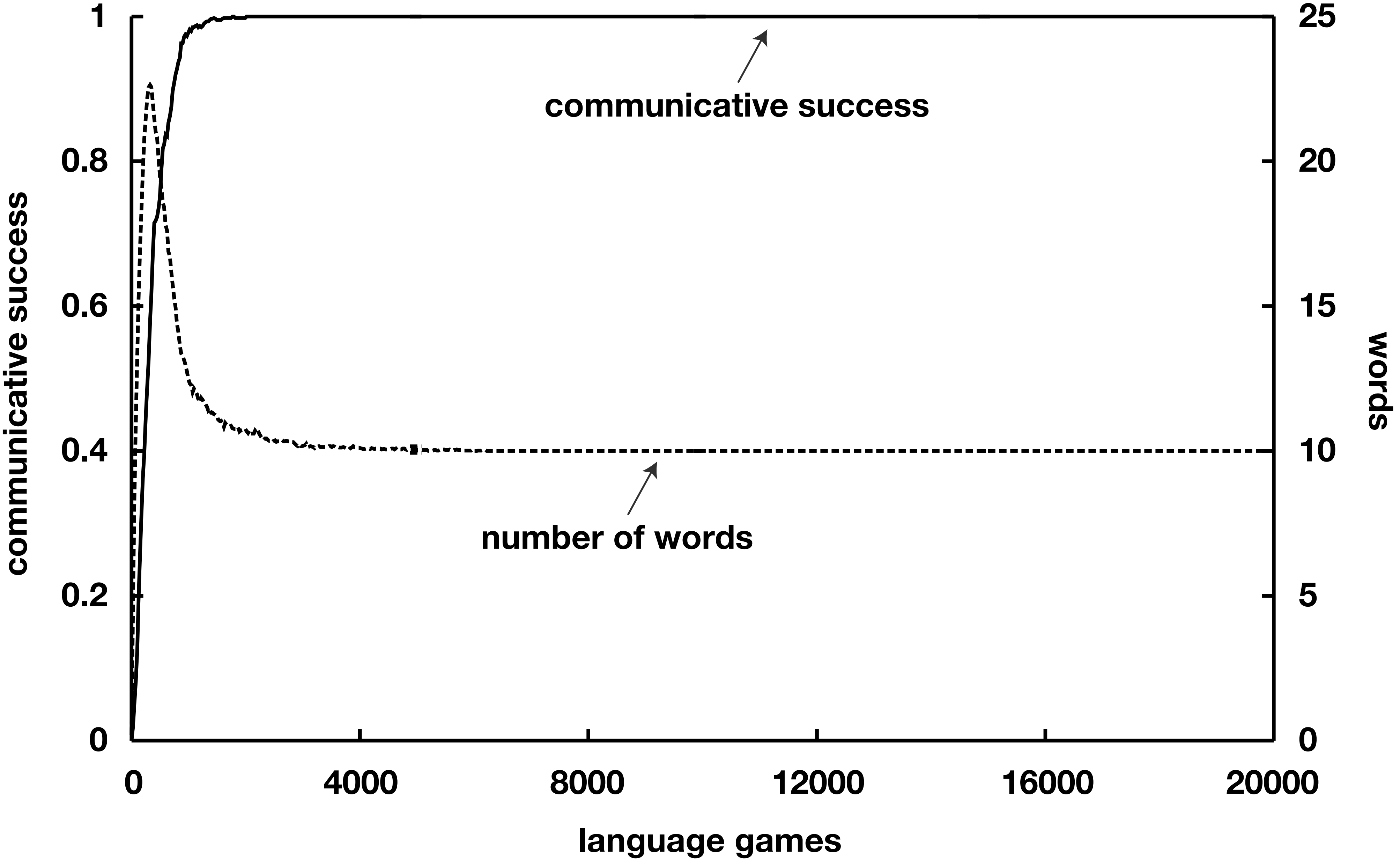
merged into the new individual i_{210} because the agent observed that these views belong to the same individual.

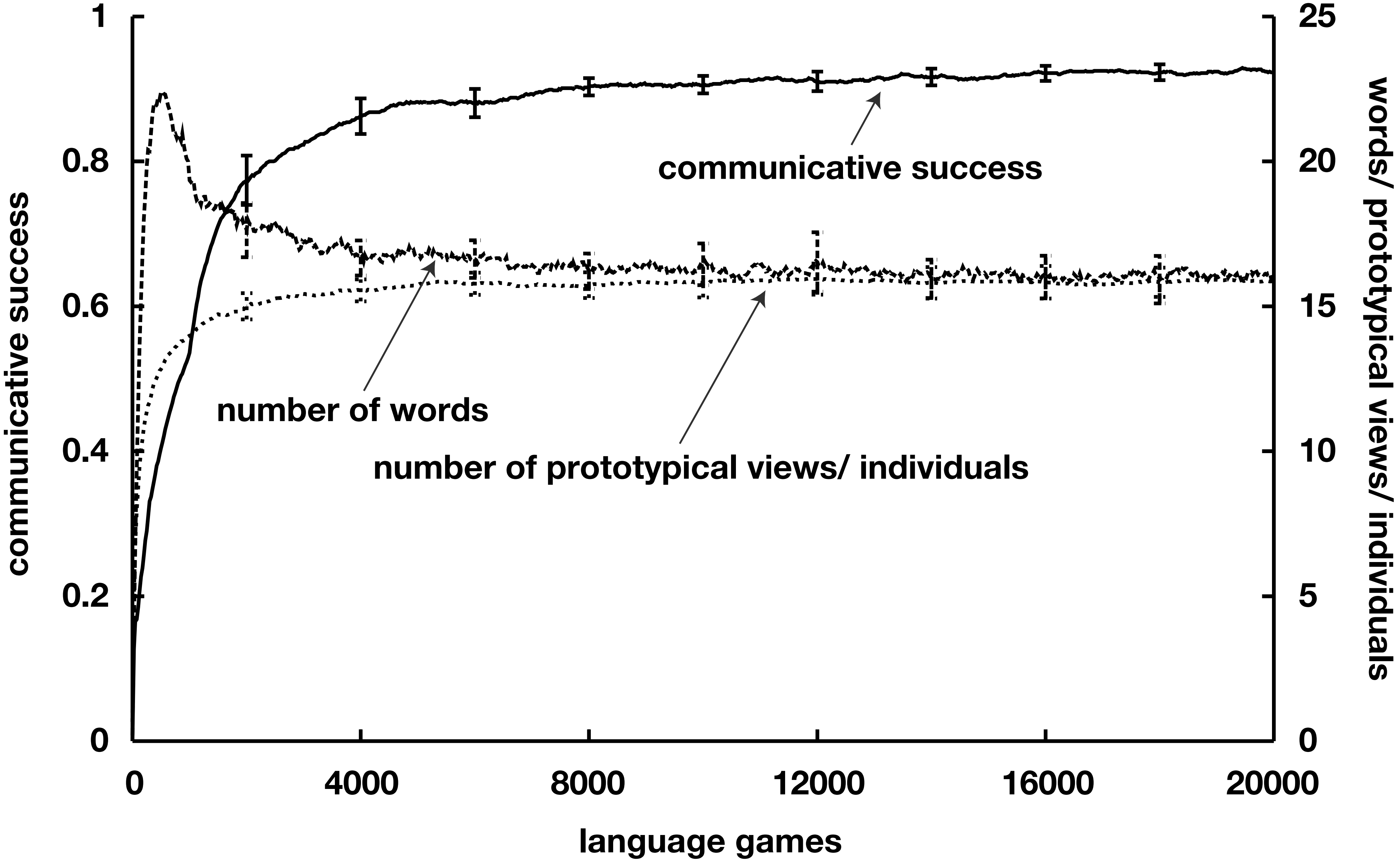
Figure 7 Inclusion of the tracking heuristic leads to a reduction of the number of individuals by about 25%, reflecting better the structure of the world. Consequently the size of vocabulary becomes smaller as well. Communicative success is lower because the small number of features used here makes it hard to distinguish individuals.

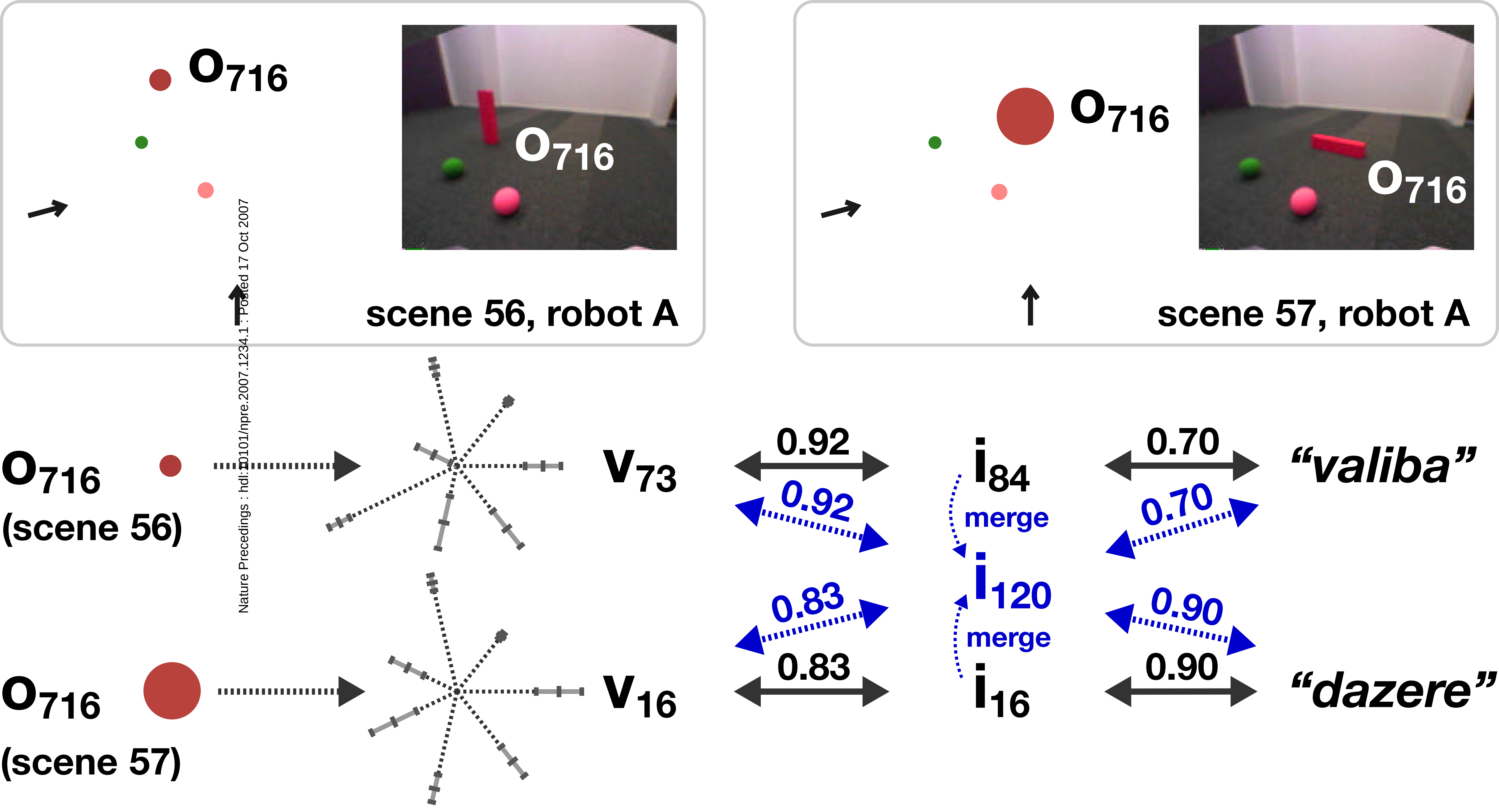

First in the third dimension

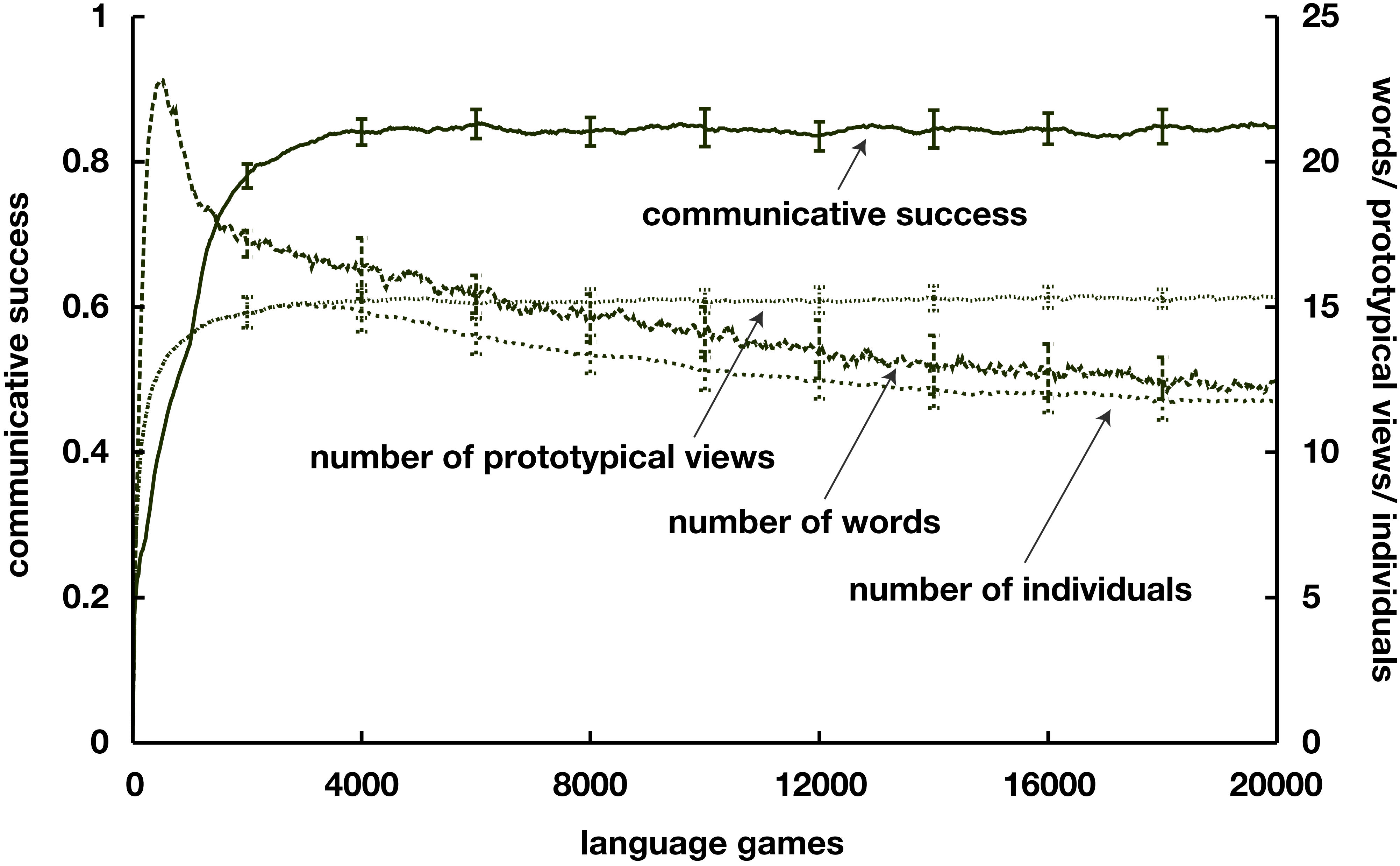

O_{716}
O_{722}
O_{708}
O_{716}
O_{722}
O_{708}
O_{716}
O_{722}
O_{708}
O_{716}
O_{716}
O_{716}
O_{716}
O_{722}
O_{708}
O_{716}
O_{722}
O_{708}
O_{722}
O_{716}
O_{708}
0
50
100
150
200
250
-50
0
50
100
150
200
250
300
0
50
100
150
200
250
300
350
400
450
0
20
40
60
80
100
120
140
160


sensory experience
prototypical views
individuals
words
o_{708}
o_{716}
o_{722}
0.21
0.08
0.13
0.17
v_{57}
v_{16}
v_{44}
v_{30}
0.57
0.83
0.72
0.89
i_{72}
i_{16}
i_{49}
1.00
0.10
0.30
0.90
0.40
1.00
“fesire”
“kaneda”
“dazere”
“doniku”
“vebewa”
association probability
distance in sensory space
legend: ● o_{708}
green/red: 0.80
luminance: 0.79
yellow/blue: 0.53
x position: 0.45
height: 0.32
width: 0.23
y position: 0.61


communicative success
number of words
communicative success
words
language games
1
0.8
0.6
0.4
0.2
0
25
20
15
10
5
0
0
4000
8000
12000
16000
20000


1
0.8
0.6
0.4
0.2
0
25
20
15
10
5
0
0
4000
8000
12000
16000
20000
communicative success
words/ prototypical views/ individuals
language games
communicative success
number of words
number of prototypical views/ individuals


o716
o716
scene 56, robot A
o716
o716
scene 57, robot A
o716
(scene 56)
v73
0.92
0.92
i84
merge
i120
merge
i16
0.70
0.70
"valiba"
o716
(scene 57)
v16
0.83
0.83
0.90
0.90
"dazere"


communicative success
words/ prototypical views/ individuals
language games
0
0.2
0.4
0.6
0.8
1
0
5
10
15
20
25
0
4000
8000
12000
16000
20000
communicative success
number of prototypical views
number of words
number of individuals