Multi-Dimensional Meanings in Lexicon Formation

Pieter Wellens and Martin Loetzsch

This paper is the author's draft and has now been officially published as:

Wellens, Pieter and Loetzsch, Martin (2012). Multi-dimensional meanings in lexicon formation. In Luc Steels (Ed.), *Experiments in Cultural Language Evolution*, 143 – 166. Amsterdam: John Benjamins.

Abstract

This chapter introduces a language game experiment for studying the formation of a shared lexicon when word meanings are not restricted to a single domain, but instead consist of any combination of perceptual features from many different domains. The main difficulty for the language users is that upon hearing a novel word they cannot be sure which aspects or properties of the referred object comprise the meaning of the word. We introduce an Adaptive Language Strategy which pays considerable attention to the adaptive nature of individual word meanings and allows the language user to use its linguistic items in a flexible manner, leading to extensive re-use. Using grounded language game experiments we show that this Adaptive Strategy elegantly copes with the problems introduced by using embodied robotic data and allows scaling towards large meaning spaces and population sizes. The strategy is further compared to a second one which lacks some of the adaptive and flexible features of the first strategy, and show that this non-adaptive strategy struggles to keep a high level of performance under taxing conditions.

1. Introduction

Previous chapters of this book have shown how a set of agents is able to bootstrap a lexicon for expressing perceptually grounded categories through evolutionary situated language games. But all case studies assumed (i) that word meanings

are concerned with a single semantic domain, e.g. colors, individuals, body postures, spatial relations, and (ii) that an utterance contains only a single word. Obviously none of these restrictions holds for human natural languages. Many words express bundles of categories and thus incorporate many different semantic domains. Moreover almost all utterances are compositional: More than one word is used to cover the set of categories that the speaker intends to convey. Human languages are therefore compositional (different words cover meaning sets that can be combined into a compositional utterance) rather than holistic (a single word covers all of the meaning).

Compositionality and multi-dimensionality are intertwined because if one word can cover many concepts and multiple words can be used, speakers need to decide how to divide up concepts over different words, and hearers are faced with the problem of finding out the part of meaning expressed by each word. For example, suppose the speaker wants to express [tall green bright block]. He may choose to do this holistically with a single word, say "dobido", that expresses all these concepts, or with two words, for example, "bado" meaning [tall green] and "zobo" [bright block]. But how can the hearer then know that "bado" does not mean [tall green bright] and "zobo" [block] or any other combination of possibilities? The problem grows exponentially, as opposed to linearly. For example, an object represented by twenty features leads to over one million possible subsets. Clearly a naive solution will not do.

The problem of compositionality and multi-dimensionality is commonly discussed in the philosophy of language literature in relation to the problem of "referential indeterminacy", as introduced by Quine (1960) through his famous "Gavagai" example. Quine tells the story of a native who points to a rabbit and says "Gavagai" to a visitor who does not know his language. How can the visitor know what the native means? Besides "rabbit" he could also mean "brown furry animal", "look at those long ears", "my pet", or "what we are going to have for dinner tonight".

The contributions of this chapter are based on new experiments for real world grounded data obtained from embodied language games with a new strategy which tries to avoid the difficulties of the competitive approaches that have been explored previously. The Talking Heads experiments (Steels et al., 2002) also used embodied data although there are several differences. The reported experiments are based on humanoid robots (see figure 1), which introduces additional problems, and the agents are given no additional cognitive mechanisms to reduce uncertainty. The language strategy thus needs to cope with high levels of uncertainty regarding word meanings. We have also conducted experiments where the number of possible di-

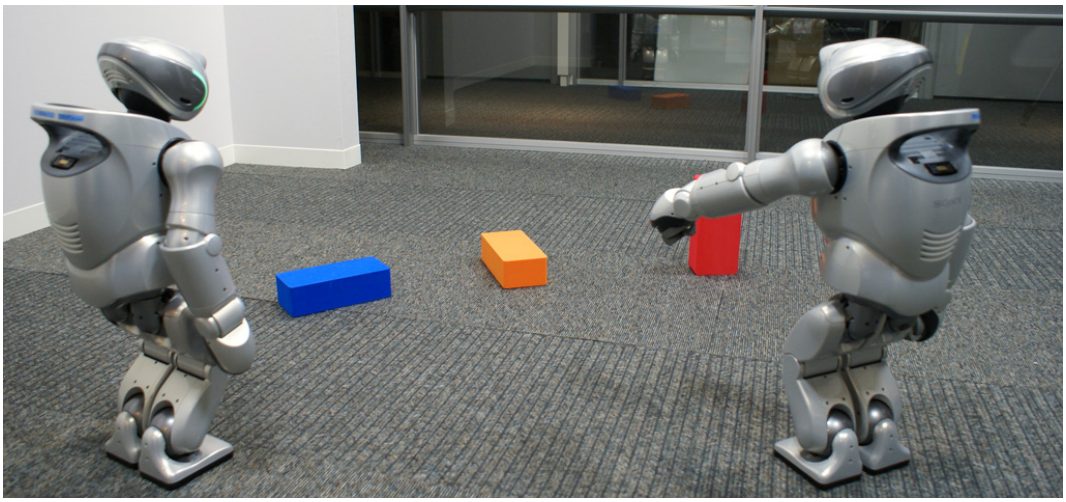


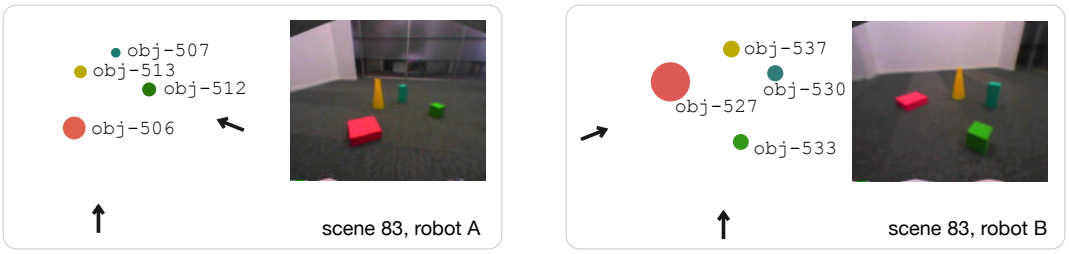
Figure 1. *Sony humanoid robots playing a language game about physical objects in a shared scene.*

mensions is significantly scaled up compared to earlier experiments. The results indicate that the methods proposed in the literature do not scale up well in the case of multi-dimensional word meanings. This has lead to our new proposal which is based on the idea that agents continuously *shape* the meaning of a word. We call this an *adaptive* strategy in contrast to the *competitive* strategies discussed in earlier work.

2. Experimental Setup

The robotic setup we use is almost identical to the one introduced in the earlier chapter describing the Grounded Naming Game (Steels & Loetzsch, 2012, this volume) and similar to many other experiments that investigate the cultural transmission of language in embodied agents (e.g. Steels, 1999, Steels & Loetzsch, 2008, see Steels, 2001 for an overview).

Just like the Grounded Naming Game the experimental setup involves two humanoid robots (Fujita et al., 2003) with the ability to perceive physical objects in a shared environment using their cameras, to track these objects persistently over time and space and to extract features from these objects. The robots must establish



obj-512	obj-507	obj-513	obj-506	obj-533	obj-530	obj-537	obj-527
width-2	*width-1*	width-2	*width-3*	width-2	*width-2*	width-2	*width-4*
height-2	*height-2*	height-4	height-2	*height-3*	*height-3*	height-4	height-2
lum-2	*lum-2*	lum-3	lum-2	lum-2	*lum-3*	lum-3	lum-2
green-red-2	green-red-1	green-red-3	green-red-4	*green-red-1*	green-red-1	green-red-3	green-red-4
yell-blue-2	yell-blue-3	yell-blue-1	*yell-blue-2*	yell-blue-2	yell-blue-3	yell-blue-1	*yell-blue-3*
x-3	x-4	x-4	*x-2*	*x-2*	x-4	x-4	*x-3*
y-1	*y-2*	*y-3*	*y-3*	*y-2*	*y-1*	*y-2*	*y-4*

Figure 2. *Visual perception of an example scene for robot A and B. On the top, the scene as seen through the cameras of the two robots and the object models constructed by the vision system are shown. The colored circles denote objects, the width of the circles represents the width of the objects and the position in the graph shows the position of the objects relative to the robot. Black arrows denote the position and orientation of the two robots. On the bottom, a subset of the features that were extracted for each object are shown. Since both robots view the scene from different positions and lighting conditions, their perceptions of the scenes, and consequently the features extracted from their object models, differ. Features that are different between the two robots are printed in italics.*

joint attention (Tomasello, 1995) in the sense that they share the same environment, locate some objects in their immediate context, and know their mutual position and direction of view. Finally, there have to be non-linguistic behaviors for signaling whether a communicative interaction was successful and, in case of failure, the robots need to be able to point to the object they were talking about.

The robots maintain continuous and persistent models about the surrounding objects using probabilistic modeling techniques (Röfer et al., 2004; Spranger, 2008). As a result, each agent has a representation of every object in the scene, including estimated position, size and color properties (see the top of Figure 2). From each such model, values on seven continuous *sensory channels* are extracted, being the position of the object in an egocentric coordinate system (`x` and `y`), the estimated

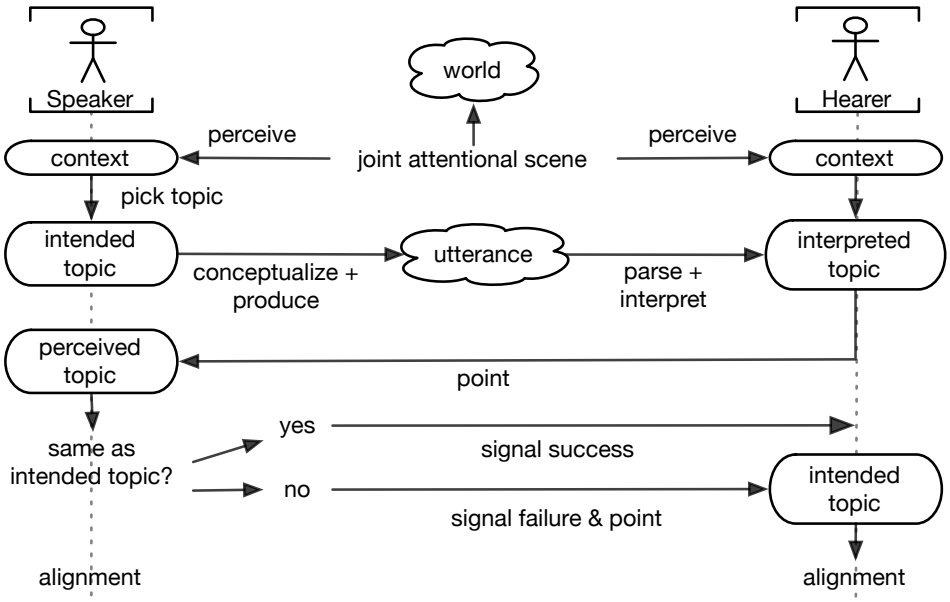


Figure 3. *Flow of one language game. A speaker and a hearer follow a routinized script. Populations of agents gradually reach consensus about the meanings of words by taking turn being speaker and hearer in thousands of such games.*

size (`width` and `height`), the average brightness (`luminance` or `lum`), average color values on a green/red and a yellow/blue dimension (`green-red` and `yell-blue`). Optionally also the uniformity of the brightness and color values within the object (as the standard deviation of all pixels within the object region in the camera image) and the maximum and minimum values of these color channels can be added, resulting in a total of sixteen continuous channels per object.

Channel values are scaled between the interval of 0 and 1, which is then split into four regions, a technique that could be compared to discrimination trees (Steels, 1997; Smith, 2001). One out of four Boolean features is assigned to an object for each channel according to the intervals of each channel value. For example the green/red value for `obj-506` in Figure 2 is `0.88`, so the assigned feature is `green-red-4`. We refer to the list of objects with their associated features as the *sensory context*.

Just like in the Naming Game all agents start with empty lexicons and have never before seen any of the physical objects in their environment. At the beginning of a game the agents establish a *joint attentional scene* (Tomasello, 1995) – a situation in which both robots attend to the same set of objects in the environment and register the position and orientation of the other robot. Once such a state is reached, the game starts. One of the agents is randomly assigned to take the role of the speaker and the other the role of the hearer. Both agents perceive a sensory context (as described above) from the joint attentional scene. The speaker randomly picks one object from his context to be the *topic* of this interaction. His communicative goal will be to describe the topic such that the hearer is able to point to it. He thus constructs an utterance, inventing new words when necessary (these mechanisms are described in detail in the following section). The hearer parses the utterance using his own lexicon and points to the object from his own perception of the scene that he believes to be most probable given the utterance. In case the hearer did not point to the correct topic, the speaker will point to the object he intended otherwise he just signals success (by nodding his head). The result is that at the end of the game the hearer knows the intended topic, but not the subset of features the speaker chose to express and certainly not how the words themselves relate to this. Finally, at the end of each interaction both agents modify their lexicons slightly based on the sensory context, the topic and the utterance (*alignment*).

Since conducting thousands of such language games with real robots would be very time-consuming and also because we wanted repeatable and controlled experiments, we recorded the perceptions of the two robots (as in Figure 2) for 150 different scenes, each containing between two and four different objects of varying position and orientation out of a set of ten physical objects. A random scene from this collection is then chosen in every language game and the two different perceptions of robots A and B are presented to the two interacting agents. In these simulations, agents point to objects by transmitting the x and y coordinates of the objects (in their own egocentric reference system). The agent receiving these coordinates can transform them into a location relative to its own position using the offset and orientation of the other robot.

3. Approaches to lexical language formation

Whether the agents can successfully bootstrap and align a lexicon depends entirely on the strategy they follow during their local interactions. Only by engaging in interactions do the agents receive (indirect) feedback whether the expansions or updates of their linguistic inventory went in the right direction. Expansion can be

either invention of new lexical entries by the speaker or adoption of novel exposures to word forms by the hearer. Alignment is achieved by updating scores associated with each lexical entry. These scores generally capture usage-based statistics, for example how many times the word was used in a successful game. These scores influence the choice of words in later language games and as such self-organising principles come into play.

Compositionality and meaning uncertainty has been studied for a decade in the literature on machine learning (Oates, 2003) and through evolutionary language games. One of the first large-scale experiments in artificial language evolution (the so called Talking Heads experiment (Steels et al., 2002)) used single-word utterances, but the problem of meaning uncertainty had to be addressed already because the topic to be identified in a game of reference had many different possible features, such as height, width, location, color or shape.

The term *Naming Game* became commonly adopted to refer to language games where there is no meaning uncertainty but there is still form uncertainty, and agents have to dampen this in order to arrive at a shared coherent lexicon, for example by lateral inhibition (figure 4a. Form uncertainty is the result of different agents inventing new words for the same meaning not knowing that a word already exists in the population.

The term *Guessing Game* has come to be used to refer to language games where both types of uncertainty happen, for example the word "tall" might be hypothesized at some point by a language user to involve the height as well as the width dimension because the topic in the language game had both characteristics (figure 4b. Meaning uncertainty arises because multiple meanings can possibly be associated to a novel word and the hearer cannot, with only one exposure, determine which meaning is intended by the speaker.

The earliest approaches to deal with meaning uncertainty (Steels, 1999) used a *discriminative approach*: The hearer tried to make the best possible guess of the meaning of an unknown word and stored this hypothesis. The best possible guess would be the one that is most discriminative in the present context or most salient from a perceptual point of view, or that had the highest score based on past usage. If in another context the hypothesized meaning did not work out, a new hypothesis was stored. Each meaning-form relation in the lexicon has a score and the scores are updated using the lateral inhibition dynamics also effective for damping synonymy: If a particular word meaning lead to a successful game, its score is increased and the score of competing meanings is decreased. If it lead to an unsuccessful game, the scored is decreased. An example of this approach has been shown in the ear-

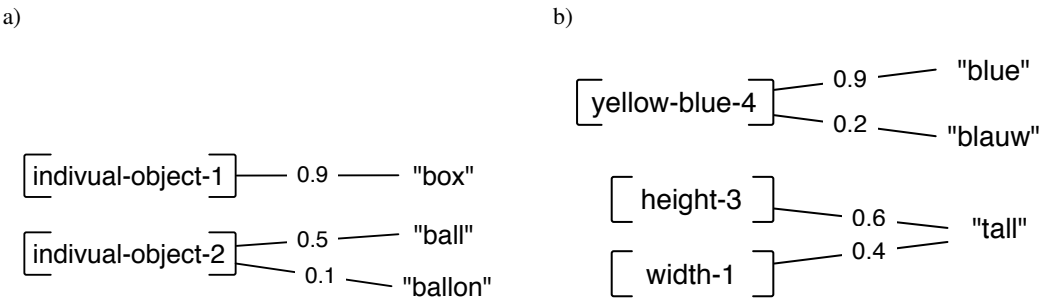


Figure 4. *Distinction between Naming Games and Guessing Games. Lexicons map meanings to forms. Meanings are denoted here with symbols in square brackets (e.g. 'height-3' refers to a range of heights) and forms with strings. In the Naming Game a meaning can have multiple forms (synonymy). In the Guessing Game the same form can have multiple meanings (meaning uncertainty) and the same meaning multiple forms (synonymy).*

lier chapter on space (Spranger, 2012, Section 4.4). If projective, proximal and absolute spatial categorizations are possible, the learner uses the one that is most discriminative as the best hypothesis for the meaning of an unknown word.

A second approach, known as *cross situational learning* (Siskind, 1996; Smith, 2005; De Beule et al., 2006; Smith et al., 2006; Vogt & Divina, 2007), has also been proposed to tackle the problem of meaning uncertainty but still assuming that words cover only a single dimension. In this approach, agents enumerate and score from the first exposure all compatible hypotheses and gradually refine this set through cross-situational statistics. For example, they might hear "dobido" for an object with the following characteristics [tall green bright block] and store, under the assumption of atomic meanings, each characteristic as a competing meaning of the word. Next they might hear "dobido" with an object that has the characteristics [tall green bright ball], in which case the score of one competitor [block] can be decreased while the other characteristics [tall], [green], and [bright] are retained. After many interactions, the mapping with the highest co-occurrence of forms and features wins over the others and becomes the dominant meaning of the word.

These approaches work well for one-dimensional meanings. The next step has been to consider Compositional Guessing Games where meanings are bundles of categories (figure 5) and utterances can have multiple words. The competitive cross-situational approach has been extended to work in this compositional context. For

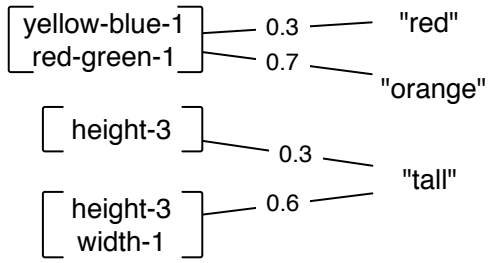


Figure 5. *Complexity increases further when there are many-to-one mappings between sets of features and words. Words can be mapped to different competing sets of categories that partially overlap each other.*

example De Beule & Bergen (2006) has shown that when there is competition between specific (many features) and general (one feature) words, the general words win over the specific ones because they are used more often – resulting again in a one-to-one mapping such as in Figure 4b. A variant of this last model will be used later to compare against the alternative language strategy that we will study in this paper.

The problem of multi-word utterances has also been studied intensely. Some proposals have argued that *holistic* utterances come first (Wray, 1998) and then (by chance) some recurring sections of utterances became associated with recurring meanings (Smith, 2008). An alternative proposal which turns out to be much more efficient and cognitively more plausible is based on *re-use* (Van Looveren, 1999): Speakers utilize all available words to cover the meaning they try to convey and if some parts are missing, new words are invented for those parts. Listeners can reconstruct partial meaning based on their own inventory and will thus have an easier time to guess the meaning of newly invented words. Alignment, as discussed in the previous sections, then does the necessary work in coordinating which words survive in the population and how much meaning they cover. Agent-based computer simulations have shown that this straightforward mechanism leads indeed to the emergence of a compositional language (Van Looveren, 1999; De Beule & Bergen, 2006; Paul, 2005), and it works both for one-dimensional meanings as for multi-dimensional ones.

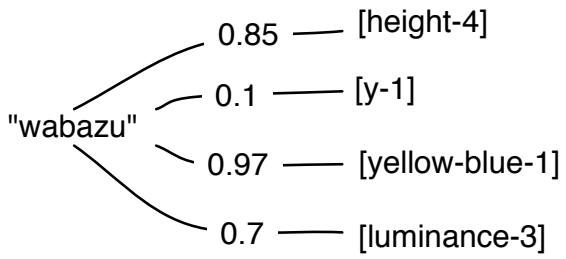


Figure 6. *A representation for a possible word "wabazu" which agents could invent. Remark that* `[y-1]` *has a score of only* 0.1 *denoting that the agent is quite uncertain whether it is part of the conventional meaning of "wabazu".*

According to the adaptive strategy, the meanings of words are multi-dimensional as they comprise any combination of perceptual features. This can be exploited by not only keeping scores related to the entire lexical entry, as is done in competition-based strategies (see Figure 4c), but by maintaining *certainty scores* for each individual meaning component associated with the word. In what follows, we refer to this representation of meaning as a *weighted set*.

By allowing the lexical entries to internally keep track of the certainty of their associated features and allowing agents to modify these scores based on use the representation becomes *adaptive*, hence the name Adaptive Strategy. Because the meaning representation itself is adaptive, the agents are, in contrast to most cross situational learning models, not required to maintain and update a set of competing meanings for a new word. It follows that during production and interpretation (discussed in detail later) there is no need to choose between competing hypotheses since there is only a single hypothesis. As an example of this representation, the meaning of the artificial word "wabazu" in Figure 6 *is* the *complete set* of scored associated features.

The rest of this paper describes first the evolutionary language game setup used in the experiments reported here. Then the adaptive strategy is presented in detail, followed by a discussion of experimental results and a comparison with the competitive learning method.

4. An Adaptive Strategy for Word Learning

4.1. Fine-Grained Adaptive Meaning Representation

The adaptive strategy maintains at all times only one meaning per word but makes the representation of this meaning more powerful and its use more flexible. The word "dobido" is at the first exposure stored with the meaning $M =$ [green tall box] and a certainty function $f_M(x) \in]0,1]$ for each characteristic x in the meaning M. In this case the features are [green], [tall] and [box]. This representation is strongly related to fuzzy set theory (Zadeh, 1965) and prototype theory (Rosch, 1973).

4.2. Flexible Re-use in Language Processing

On the processing side a strategy for the compositional guessing game needs to specify the following *processing* components:

Conceptualization and production: Which subset of the features of the topic should the speaker express and which combination of words does he use to express them?

Parsing and interpretation: How should the hearer reconstruct the meaning of the received utterance and which object should he point to?

Invention and adoption: Which diagnostics and repair strategies do the agents have? How can they invent and adopt new words?

Alignment: How do the agents update their inventories?

4.2.1. *Conceptualization and Production*

How can the speaker maximize communicative success when expressing the topic? Expressing all perceptual features of the topic is not the best strategy because not all features are equally discriminative and the speaker might not (yet) have the expressive power to express every feature of the topic. For example when besides the topic other objects also are [red] it might be better not to express [red]. A good conceptualization for the speaker is thus a subset of features that is both discriminative and expressible.

The adaptive strategy puts re-use of existing words at the center of its linguistic processing. Agent can re-use a word even when some characteristics of the meaning mismatch with the topic. For example if the topic has characteristics [green

small box] an agent is still allowed to use the word "dobido" which has associated meaning [green tall box]. An agent can rank his words for a given set of features by employing a *similarity function*.

The similarity function takes as input a weighted meaning M and the characteristics of an object O (e.g. the topic) and returns a number in the interval $[-1,1]$ as follows:

$$\text{Similarity}(M,O) = \frac{|M \cap_w O|_w - |M \setminus_w O|_w}{|O|} \tag{1}$$

Meaning M is a weighted set which means that for every element x in M function $f_M(x)$ keeps a certainty score. The set operations intersection $\cap_w$, set difference $\setminus_w$ and cardinality $||_w$ as used in the above similarity measure are adapted to take the certainty scores into account by using their fuzzy set counterparts. As such the intersection of two sets weighted A and B is a third weighted set C such that $C = A \cap B$ with $f_C(x) = \text{Min}[f_A(x), f_B(x)]$. Likewise the difference of two weighted sets A and B is a third weighted set C such that $C = A \setminus B$ with $f_C(x) = f_A(x)$. And finally the cardinality $|A|_w$, instead of counting the elements in A, takes the sum of the certainty scores of the elements in A, $|A|_w = \sum_{x \in A} f_A(x)$. For object O which is not a weighted set we assume that $f_O(x) = 1, \forall x \in O$.

Given a meaning M and an object O, `Similarity` takes all shared features and disjoint features between M and O. Sharing features is thus beneficial for the similarity while the opposite is true for features that are not shared. `Similarity(M,O)` will be high when M and O share many features with high certainty scores. Correspondingly, the result will be low when M and O have many disjoint features with high certainty scores. Furthermore when M is a subset of O the `Similarity` can never be negative because $|M \setminus_w O|_w = 0$. Negative similarity means that the weight of the disjoint features is larger than that of the shared features. Some examples:

`Similarity(((a 1) (b .5) (c .7)), (a b c))` $= \frac{2.2 - 0}{3} = 0.73$
`Similarity(((a .5) (b .9) (c .3)), (d e f))` $= \frac{0 - 1.7}{3} = -0.57$
`Similarity(((a .9)), (a b c))` $= \frac{0.9 - 0}{3} = 0.3$
`Similarity(((a .5) (b .5) (c .5)), (a c d))` $= \frac{1 - 0.5}{3} = 0.17$
`Similarity(((a .9) (b .1) (c 1)), (a c d))` $= \frac{1.9 - 0.1}{3} = 0.6$

Finally we extend the Similarity function so that given an utterance U and an object O, Similarity(U,O) = *Similarity*$(\bigcup_{w \in U} \text{Meaning(w)}, O)$ where the union $\bigcup$ takes the fuzzy union between sets A and B such that $C = A \cup_w B = A \cup B$ with $f_C(x) = \text{Max}[f_A(x), f_B(x)]$.

In production the speaker wishes to *maximize* the similarity between his utterance and the topic and *minimize* the similarity between that same utterance and the other objects in the context. Production then is a search process which, given a context C and a topic T, finds the utterance as follows:

$$\text{Produce(C,T)} = U \text{ which maximizes Similarity(U,T)} - \underset{O \in C \setminus t}{\text{argmax}}[Similarity(U,O)].$$

The final utterance is that subset of the lexicon that strikes the optimal balance between being most similar to the topic and being most distant from the other objects of the context. This results in a context sensitive multi-word utterance and involves an implicit, on-the-fly discrimination using the lexicon.

The hearer points to the object O which has highest similarity with the utterance according to his lexicon:

$$\text{Interpret(C,U)} = O \text{ which maximizes Similarity(U,0)}$$

The most important benefit of using a similarity measure is the great flexibility in word combination, especially in the beginning when the features have low certainty scores. The flexibility allows the agents to use (combinations of) words that do not fully conform to the meaning to be expressed, resembling what Langacker (2002) calls *extension*. The ability to use linguistic items beyond their specification is a necessity in high dimensional spaces for maintaining a balance between lexicon size and coverage (expressiveness).

4.3. Invention and Adoption

The flow of the language game depicted in Figure 3 does not include learning. Both production and parsing can fail, forcing the speaker or hearer to extend their lexicons. In addition, even when production and parsing were successful the game itself might be a failure (e.g. because of misaligned lexicons), again forcing the agents not only to adapt their scores but also to alter their linguistic inventories more drastically.

In production, before actually uttering the utterance the speaker first places himself in the role of the hearer and tries to interpret the utterance himself, a process called *re-entrance* (Steels, 2003). Re-entrance rests on the same principles as the *Obverter* strategy introduced by Oliphant & Batali (1997) and later also incorporated in the models of Kirby (2002). When re-entrance leads the speaker to a different topic than his own intended topic it means that no utterance could be found

which successfully refers to the topic in the given context. In this case the speaker invents a new form (a random string) and associates to it, with very low initial certainty scores, all features of the topic that were not yet expressed in the utterance. For example if the topic consists of categories [tall blue distant box] and the agent came up with utterance "wabado dipozu" with a combined meaning of [tall black distant] then, in case re-entrance fails, the speaker will invent a word with initial meaning [blue box]. More formally, given an utterance U and a topic T the new meaning M_{new} is arrived at as follows:

$$M_{new} = T \setminus \bigcup_{w \in U} \text{Meaning(w) with } f_{M_{new}}(x) = 0.05 \tag{2}$$

When the hearer encounters a novel word in the utterance he needs a way to associate an initial representation of meaning with this novel word form. The hearer first interprets the words he does know and tries to play the game without adopting the novel forms. At the end of the game, when he has knowledge of the topic (see Figure 3), the hearer associates all unexpressed features with the the novel form, which is the exact same as the speaker does [see equation (2)]. Just as with invention the initial certainty scores for the features start out very low, capturing the uncertainty of this initial representation. Although invention and adoption use the same strategy to come up with the initial hypothesis the hearer uses his own interpretation for this and in practice the speakers and hearers initial hypotheses can be quite distinct.

With invention and adoption covered this leaves us with the alignment strategy. From the perspective of self-organizing dynamics, alignment is the most crucial part of the language game.

4.4. Alignment

In the previous sections we have seen how the similarity measure allows the agents to flexibly use their words. This flexibility, which leverages the power of re-use, entails that the agents can express and interpret meanings of which some categories mismatch with the topic. For example for a topic composed of categories [blue small ball] and an utterance consisting of the words "baduzo" [blue] and "fudega" [distant ball] the categories [blue] and [ball] match whereas [distant] does not. In such a case Langacker (2002) calls the use of the word "fudega" an extension which entails "strain". This strain in usage in turn affects the meanings of these linguistic items. The certainty score of the features that raised similarity are

incremented (rewarded) and the others are decremented (punished). This resembles the psychological phenomena of entrenchment, and its counterpart semantic erosion (also referred to as semantic bleaching or desemantisation). In the example [blue] associated to "baduzo" and [ball] associated to "fudega" are entrenched while the association of [distant] with "fudega" gets eroded. If by decrementing the certainty score of a feature, it drops below zero, the feature is removed from the word meaning, resulting in a more general word meaning.

Features cannot only erode they can also be added when an agent believes his current word meaning is not similar to the conventionalized one. There is only one case in which an agent can reasonably believe this to be the case, namely when he, as a hearer in a language game knows all the spoken words but yet failed at pointing to the intended topic. Obviously this means that the speaker must have different features associated with the spoken words. In this case the hearer adds all unexpressed features of the topic, again with very low certainty scores, to *all* uttered words, thus making the meanings of those words more specific.

Combining similarity-based flexibility with entrenchment and semantic erosion, word meanings gradually shape themselves to better conform with future use. Repeated over thousands of language games, the word meanings progressively refine and shift, capturing frequently co-occurring features (clusters) in the world, thus implementing a search through the enormous hypothesis space, and capturing only what is functionally relevant.

5. Experimental Results

We tested the Adaptive Strategy by letting populations of 25 agents play ten repeated series of 50000 language games. Since only two agents participate in each game, a single agent takes part in only 4000 out of the total 50000 games. Two measures, being communicative success and average lexical coherence, are shown in Figure 7. The agents register communicative success when the hearer points to the topic the speaker intended. It is a Boolean measure which is averaged over the last 500 interactions. Communicative success is a usage-based measure showing whether the agents are capable of bootstrapping and aligning a language system suitable for playing the language games. High communicative success does not necessarily require or entail high lexical coherence. Lexical coherence shows how aligned (or coherent) the lexicons of the agents are, by comparing the similarity of the meanings for the same word form in all agents. Coherence is high when agents associate the same set of features to the same word forms, it is low when meanings for the same word forms differ significantly among agents.

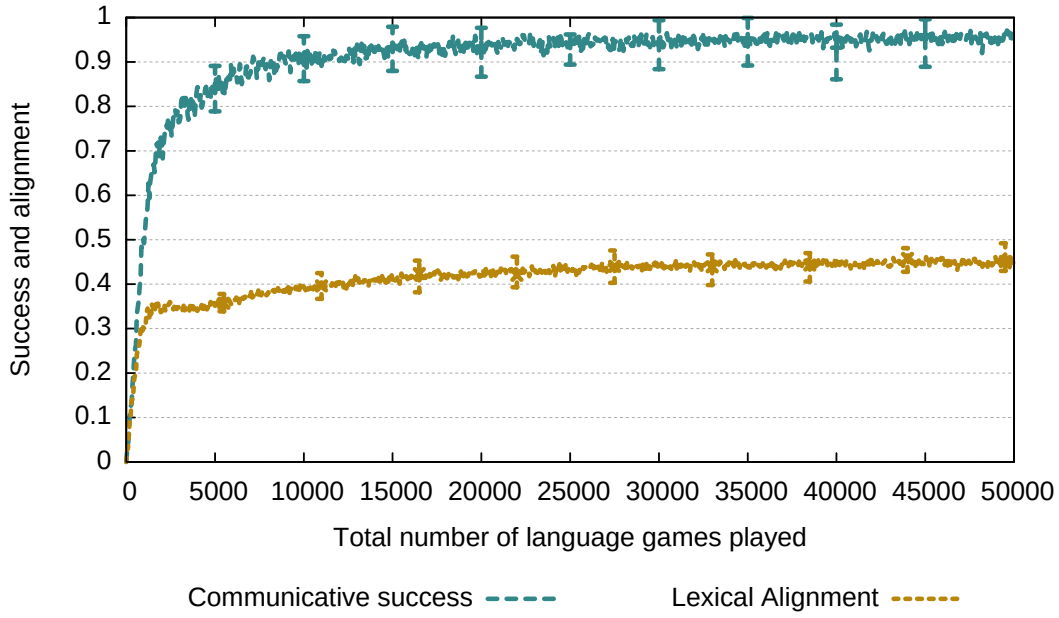


Figure 7. *Dynamics using the Adaptive Strategy as explained in Section 4 for a population of 25 agents averaged over ten runs of 50000 interactions. After each interaction (plotted on the x-axis) the communicative success and the average lexical coherence is plotted. Error bars repesent the minimum and maximum across ten different experimental runs.*

From quite early on (at around interaction 10000), the agents communicate successfully in more than 90% of the cases. Due to the alignment dynamics, coherence increases but nevertheless stagnates around 0.5. This means that the average similarity over the population for the meanings associated to the same word form is 0.5. Remember that the similarity measures ranges between -1 and 1. This result shows that the agents do not require full alignment in order to communicate successfully. One of the main reasons why alignment stops increasing is because the agents have different perceptions of the same scenes, using only the context it is thus impossible to arrive at a fully aligned communication system.

In order to understand the alignment process better, we looked at how the meaning of a single word in one agent evolves over time. The alignment dynamics continuously adapt the certainty scores of the associated features and as such shape

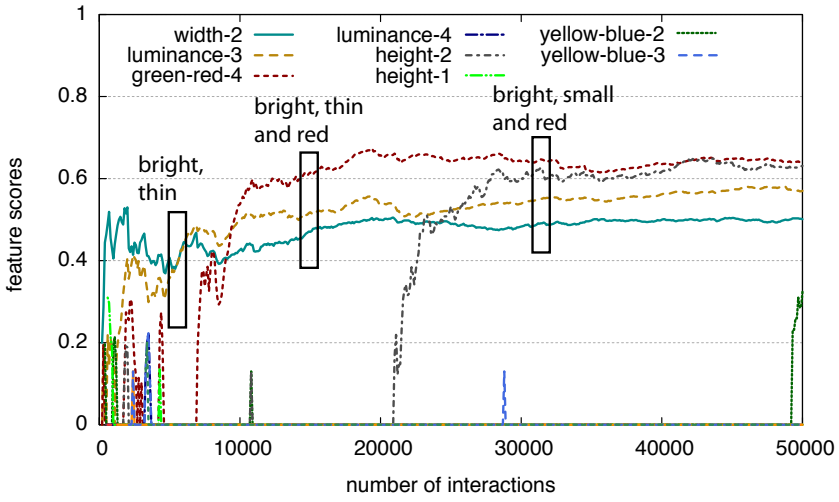
the meaning as a whole over the course of many usage events. Figure 8 gives two examples of the shaping of word meanings over time. Despite some other associations that disappear very quickly, the word in Figure 8a is initially only connected to `width-2`. Over the course of many interactions, more and more features are associated (`luminance-3` at around interaction 3000, `green-red-4` at interaction 7000 and finally `height-2` at interaction 22000). So this word changed from being very general ("thin") to very specific ("thin, low, bright and red"). The word in Figure 8b is an example of the opposite. It starts out very specific, with connections to `green-red-4`, `yellow-blue-2`, `height-2`, `width-2`, `luminance-3` ("orange, small and bright"). It loses most of these features, becoming very general ("orange") towards the end.

6. A Comparison with a Competitive Strategy

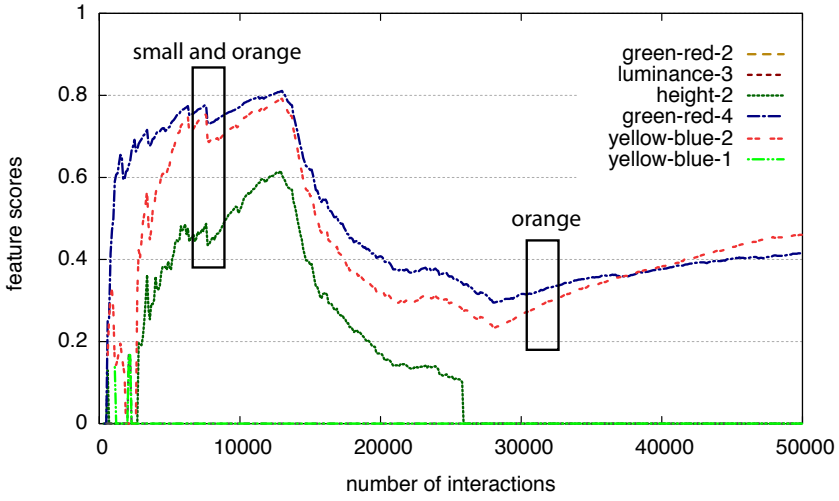
So far we have established that agents endowed with the Adaptive Strategy can successfully bootstrap and align a shared lexicon under demanding conditions, such as embodied data and different perceptions. The question remains whether the two main aspects of this strategy, namely an adaptive representation and flexible processing, are in fact responsible for this level of performance.

To investigate this we implemented a second strategy, which is inspired by the cross-situational models presented in Siskind (1996); Smith (2005); De Beule et al. (2006); Smith et al. (2006); Vogt & Divina (2007); Van Looveren (1999). Just as the model underlying the Adaptive Strategy, these models were proposed to deal with the problem of meaning uncertainty. They were however not developed to handle a communicative task as demanding as in the current experimental setup. For example most of them (Siskind (1996); Smith (2005); Smith et al. (2006); Vogt & Divina (2007); Van Looveren (1999)) assume a one-to-one mapping of a feature to a word form and thus did not allow combinations of features (see Figure 4b). In contrast to an adaptive meaning representation (i.e. every feature is scored) these models represent the uncertainty by enumerating competing hypotheses and relying on dynamics at the level of the lexicon to eliminate this competition (and thus the uncertainty) over time. Furthermore none of these models employ a type of flexible processing since they rely on a strict covering algorithm instead of similarity as in the Adaptive Strategy. In what follows strategies inspired by these models will be called Competitive Strategies, since they rely on an enumeration of competing hypotheses.

As we did for the Adaptive Strategy we briefly give an overview of the representational and processing aspects of the implemented Competitive Strategy. On



(a) General to specific



(b) Specific to general

Figure 8. *Adaptive word meanings. Each graph shows, for one particular word in the lexicon of one agent, the certainty score of the associated features. In order to keep the graphs readable, the agents have been given access to only 5 sensory channels (*`width`*,* `height`*,* `luminance`*,* `green-red`*,* `yellow-blue`*).*

the representational side the agents keep a certainty score for every word meaning association. This score reflects the amount of times the meaning was compatible with (i.e. a subset of) the topic.

From a processing perspective the Competitive Strategy works as follows.

Conceptualization and Production: A speaker can enumerate all minimal discriminative subsets of the topic and choose the discrimination that can be best covered with the words from his lexicon. A lexical entry "covers" a part of the meaning when its associated features are a subset of the features of the meaning and is thus stricter than using a similarity measure. The end result is that the speaker will find the most successful (based on the averaged certainty scores) minimal utterance to describe the topic.

Parsing and Interpretation: Upon hearing the utterance the hearer generates all possible interpretations for the words in the utterance, by taking the unions of their meanings. He ranks these interpretations based on the average certainty scores of the words and compares these to each object in the context. Finally he points to that object which is compatible with the highest ranked interpretation.

Invention and Adoption: Even when the speaker cannot fully cover any minimal discrimination of the topic, denoting he cannot construct an utterance, he is still able to find the best partial coverage. Using this partial coverage or utterance he invents a new word associating all uncovered features of the partially covered discrimination to it. In adoption the hearer does more or less the same when he encounters a novel word form. Using the feedback from pointing he knows the intended topic and finds the best partial interpretation for it based on the words he knows. He then creates a new word for the novel form and associates all uncovered attributes to it.

Alignment: The speaker updates the certainty scores of his used words. He also dampens synonyms based on lateral inhibition dynamics as used in the Naming Game (Steels & Loetzsch, 2012, this volume). The hearer has a slightly more refined updating mechanism. Since he now knows the intended topic and the utterance he increases the certainty score for all lexical entries that could have been part of a successful interpretation. Furthermore if the game is a failure and there was no possible correct interpretation given his lexicon he will generate meaning competitors (i.e. words with the same form

but different meaning) for the words in the utterance such that in the future a successful interpretation would be possible.

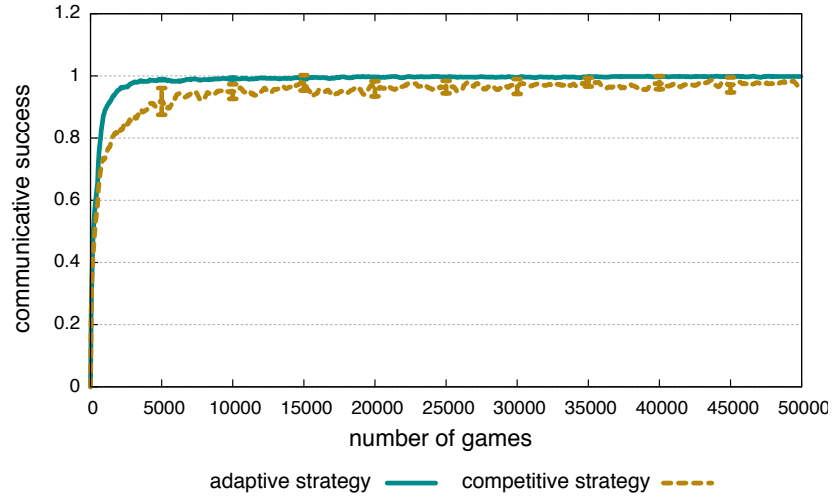
6.1. Experimental results

We compared both strategies by endowing one population of agents with the Adaptive Strategy and one with the Competitive Strategy. However, the Competitive Strategy is not capable of reaching communicative success under the same experimental conditions as those for the Adaptive Strategy shown in Figure 7. We thus reduced the population size from 25 to 5, the amount of perceptual features per object from 16 to 5 (only width, height and three color features remain) and also allowed both agents (robots) to share the same perception of the scene. In other words they used only one camera and thus perceived the exact same perceptual features per object.

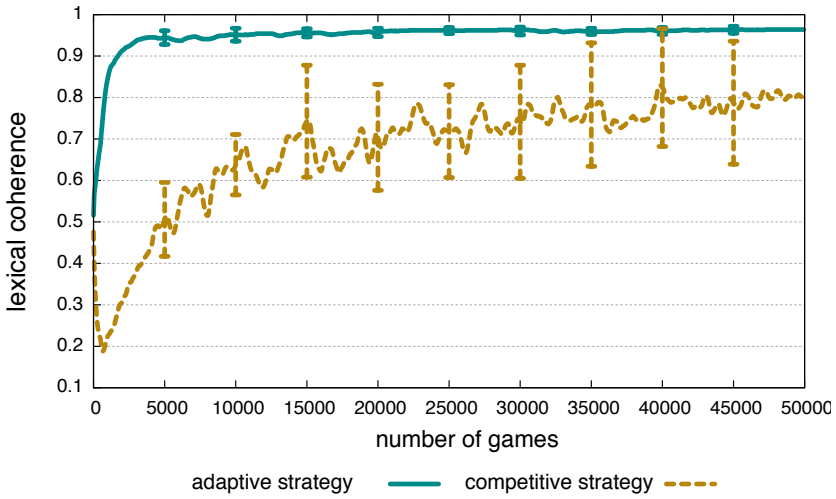
We again measured communicative success and lexical coherence as shown in Figure 9. Even though the above limitations were in place, the Adaptive Strategy reaches higher levels of communicative success, reaches them faster and they remain more stable. The main reason for this can be found in graph (b) which show that lexical coherence grows much more slowly in the population with the Competitive Strategy.The large amount of meaning competitors in the lexicon requires a long time to get successfully eliminated.

In these experimental runs both the population size and the number of perceptual features were scaled down and the problems of different perceptions were eliminated by sharing one view of the scene. Starting from the relaxed experimental setup as used for Figure 9 we again measured communicative success but altered one of the three parameters so that the effect of the parameter could be shown in isolation. Results are shown in Figure 10. The bars in the chart depict average communicative success of the final 500 games. The total number of games played was 100000 for graph (a) and 50000 for graph (b) and (c). Graph (a) shows the impact of scaling up the size of the population from five to one hundred. Although the Competitive Strategy works fine for five agents it has difficulties to cope with an increase in the number of agents. More agents lead to more inventions and thus tax the alignment much heavier.

Next the number of perceptual features per object was increased form 5 to 16, starting with `width, height, luminance, green-red, blue-yellow` and adding `x y`, standard deviations for the three color channels and finally adding a maximum and minimum value for each color channel giving a total of 16 features.

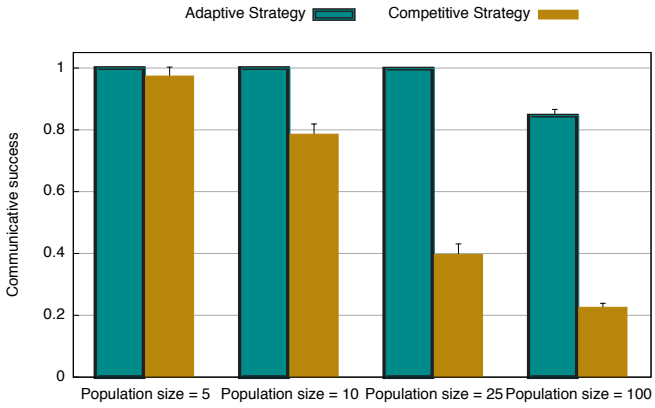


(a) Communicative Success

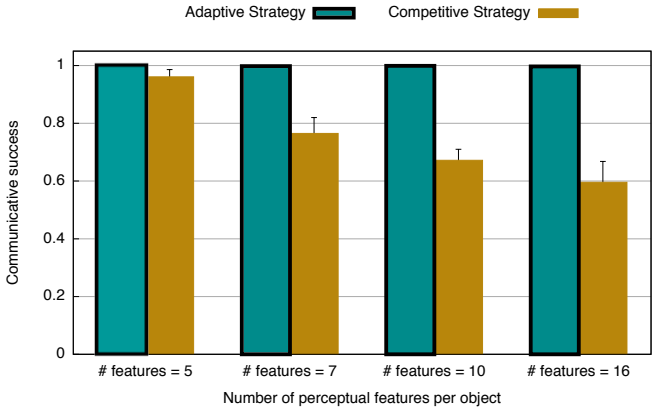


(b) Lexical Coherence

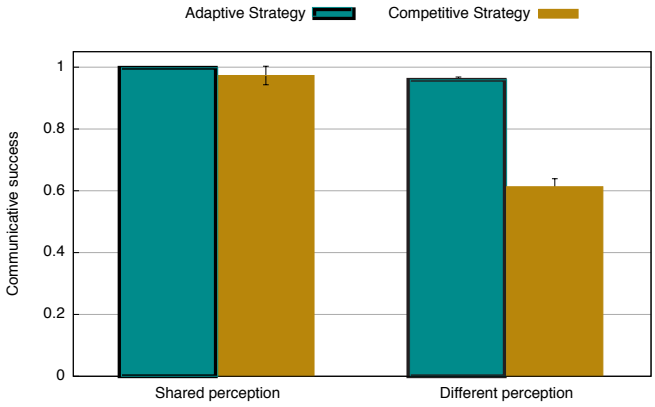
Figure 9. *Comparison of communicative success (graph a) and lexical coherence (graph b) for the Adaptive and Selectionist Strategy using a population of only five agents averaged over ten runs of 50000 interactions. Furthermore the agents share the same perception (camera) of the scene and are thus bypassing the difficulties introduced by differences in perception. Also the number of features per object was reduced from ten to five, keeping only the five most informative features (width, height and three color features).*



(a) Population size



(b) Number of perceptual features



(c) Perception

Figure 10. *Comparison of robustness and scaling of the Adaptive and Selectionist Strategy for a population of five agents averaged over ten runs after 100000 interactions for the graph a) and 50000 interactions for graphs b) and c).*

Again the Competitive Strategy struggles with the increasing number of features. This is mainly due to the exponential increase of potential meaning competitors as the number of features increases. Finally graph (c) shows that the Competitive Strategy is not robust to different perceptions of the scene for which the main reason lies in the discriminatory phase and the non-flexible application of the lexical entries.

7. Conclusion

Although word learning is often equated to a mapping task in which the learner needs to map word forms to meanings or concepts, we investigate a different approach that sees the child as constructing and gradually *shaping* word meanings Bowerman & Choi (2001). The hypothesis is that "... the use of words in repeated discourse interactions in which different perspectives are explicitly contrasted and shared, provide the raw material out of which the children of all cultures construct the flexible and multi-perspectival – perhaps even dialogical – cognitive representations that give human cognition much of its awesome and unique power" (Tomasello, 1999, p. 163). Children cannot have at hand all the concepts and perspectives that are embodied in the words of the language they are learning – they have to construct them over time through language use. Moreover, the enormous diversity found in human natural languages (Haspelmath et al., 2005; Levinson, 2001) and the subtleties in word use (Fillmore, 1977) suggest that language learners can make few a priori assumptions and even if they could, they still face a towering uncertainty in identifying the more subtle aspects of word meaning and use. The adaptive representation can be seen as an implementation of the notion of a prototype as introduced by Rosch (1973). Also more recently (Tomasello, 2001, p. 133) indirectly criticized the competitive approach as follows "... One approach to the problem of referential indeterminacy in the study of lexical acquisition is the so-called "constraints" approach. In this view a learner, [...], attempts to acquire a new word by: (1) creating a list of possible hypotheses about how the new word "maps" onto the real world, and (2) eliminating incorrect hypotheses in a semi-scientific manner. [...]. The problem is that, [...], there are simply too many possible hypotheses to be tested in a given case.".

We believe the core ideas of the presented model, namely adaptive representation and flexible processing, can also be extended to more abstract grammatical constructions for example in the formation of semantic categories. The core ideas can be seen as simplifications of the mechanisms at work in grammaticalization (Heine, 1997), with adaptive representations referring to mechanisms of erosion

and desemantisation and flexible processing to more subtle cognitive capabilities of extension and metaphor. To pursue these ideas would be a very interesting direction for further investigation. Another interesting avenue of research is to combine the strengths of both the Adaptive and Competitive Strategy. Especially for the emergence of syntactic categories the Competitive Strategy might bring an added value or even be necessary, as for example in Gong et al. (2009).

We introduced a fully implemented model for word learning that can cope with the problem of meaning uncertainty in a robust way. The key to this robustness lies in the adaptive representation of meaning and the flexible use of the linguistic items. This flexibility not only allows more pervasive re-use but also provides the required feedback to slightly adapt the representation of meaning in alignment. The model scales well with population size and features per object. It can also handle problems stemming from embodied data such as differing perceptions of the same scene. These scaling and robustness characteristics were not found when tested with another less adaptive model.

Acknowledgments

This research was carried out at the Artificial Intelligence Laboratory of the Vrije Universiteit Brussel and the Sony Computer Science Laboratory in Paris and Tokyo with additional support from FWOAL328, the EU-FET ECAgents project (IST-2003 1940), ALEAR and EUCOG II. Pieter Wellens also received funding as a research assistant at the VUB Computer Science Department. The authors are grateful to Masahiro Fujita and Hideki Shimomura of the Intelligent Systems Research Labs at Sony Corp, Tokyo for graciously making it possible to use the QRIO robots for the reported experiments. We thank Tao Gong (Max Planck Institute for Evolutionary Anthropology Leipzig, Germany) for his insightful review of this paper. We want to credit our colleagues, especially Joachim De Beule, for the many constructive discussions and Michael Spranger for his indispensable help with the robotic set-up. Finally we wish to thank our supervisor, Luc Steels, for his invaluable feedback and support. Reported experiments have been implemented within with the Babel 2 framework.

References

Bowerman, Melissa, Soonja Choi (2001). Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. In

Melissa Bowerman, Stephen C. Levinson (Eds.), *Language Acquisition and Conceptual Development*, 132–158. Cambridge: Cambridge University Press.

De Beule, Joachim, Benjamin Bergen (2006). On the emergence of compositionality. In Angelo Cangelosi, Andrew Smith, Kenny Smith (Eds.), *Proceedings of the 6th International Conference on the Evolution of Language*, 35–42. London: World Scientific Publishing.

De Beule, Joachim, Bart De Vylder, Tony Belpaeme (2006). A cross-situational learning algorithm for damping homonymy in the guessing game. In Luis Mateus Rocha, Larry Yaeger, Mark Bedau, Dario Floreano, Robert Goldstone, Allessandro Vespignani (Eds.), *Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*, 466–472. Cambridge, MA: MIT Press.

Fillmore, Charles (1977). Scenes-and-frames semantics. In Antonio Zampolli (Ed.), *Linguistic structures processing*, 55–81. Amsterdam: North Holland Publishing.

Fujita, Masahiro, Yoshihiro Kuroki, Tatsuzo Ishida, Toshi Doi (2003). Autonomous behavior control architecture of entertainment humanoid robot sdr-4x. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)*, 960–967. Las Vegas, Nevada.

Gong, Tao, James Minett, William Wang (2009). A simulation study on word order bias. *Interaction Studies*, 10, 51–76.

Haspelmath, Martin, Matthew Dryer, David Gil, Bernard Comrie (Eds.) (2005). *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Heine, Bernd (1997). *The Cognitive Foundations of Grammar*. Oxford: Oxford University Press.

Kirby, Simon (2002). Natural language from artificial life. *Artificial Life*, 8(2), 185–215.

Langacker, Ronald (2002). A dynamic usage-based model. In Michael Barlow, Suzanne Kemmer (Eds.), *Usage-Based Models of Language*, 1–63. Chicago: Chicago University Press.

Levinson, Stephen (2001). Language and mind: Let's get the issues straight! In Melissa Bowerman, Stephen C. Levinson (Eds.), *Language Acquisition and Conceptual Development*, 25–46. Cambridge: Cambridge University Press.

Oates, Tim (2003). Grounding word meanings in sensor data: Dealing with referential uncertainty. In *Proceedings of the HLTNAACL 2003 workshop on Learning word meaning from nonlinguistic data.*, 62–69. Association for Computational Linguistics.

Oliphant, Michael, John Batali (1997). Learning and the emergence of coordinated communication. *The newsletter of the Center for Research in Language*, 11(1).

Paul, Vogt (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, 167(1-2), 206–242.

Quine, Willard (1960). *Word and Object*. Cambridge, MA: MIT Press.

Röfer, Thomas, Ingo Dahm, Uwe Düffert, Jan Hoffmann, Matthias Jüngel, Martin Kallnik, Martin Lötzsch, Max Risler, Max Stelzer, Jens Ziegler (2004). Germanteam 2003. In Daniel Polani, Brett Browning, Andrea Bonarini (Eds.), *RoboCup 2003: Robot Soccer World Cup VII*, *Lecture Notes in Artificial Intelligence*, vol. 3020. Padova, Italy: Springer. More detailed in GermanTeam RoboCup 2003. Technical Report (199 pages, http://www.germanteam.org/GT2003.pdf).

Rosch, Eleanor (1973). Natural categories. *Cognitive Psychology*, 7, 573–605.

Siskind, Jeffrey (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1), 39–91.

Smith, Andrew (2001). Establishing communication systems without explicit meaning transmission. In *Advances in Artificial Life Proceedings of the Sixth European Conference, ECAL 2001*, *Lecture Notes in Artificial Intelligence*, vol. 2159, 381–390. Berlin: Springer Verlag.

Smith, Andrew (2005). The inferential transmission of language. *Adaptive Behavior*, 13(4), 311–324.

Smith, Kenny (2008). Is a holistic protolanguage a plausible precursor to language? a test case for a modern evolutionary linguistics. *Interaction Studies*, 9(1), 1–17.

Smith, Kenny, Andrew Smith, Richard Blythe, Paul Vogt (2006). Cross-situational learning: a mathematical approach. In P. Vogt, Y. Sugita, E. Tuci, C. Nehaniv (Eds.), *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication*, 31–44. Springer Berlin/Heidelberg.

Spranger, Michael (2008). *World Models for Grounded Language Games*. German diplom thesis, Humboldt-Universität zu Berlin.

Spranger, Michael (2012). The co-evolution of basic spatial terms and categories. In Luc Steels (Ed.), *Experiments in Cultural Language Evolution*. Amsterdam: John Benjamins.

Steels, Luc (1997). The origins of ontologies and communication conventions in multi-agent systems. *Journal of Agents and Multi-Agent Systems*, 1(1), 169–194.

Steels, Luc (1999). Situated grounded word semantics. In Thomas Dean (Ed.), *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99)*, 862–867. Stockholm, Sweden: Morgan Kaufmann.

Steels, Luc (2001). Language games for autonomous robots. *IEEE Intelligent Systems*, 16–22.

Steels, Luc (2003). Language re-entrance and the inner voice. *Journal of Consciousness Studies*, 10(4-5), 173–185.

Steels, Luc, Frédéric Kaplan, Angus McIntyre, Joris Van Looveren (2002). Crucial factors in the origins of word-meaning. In Alison Wray (Ed.), *The Transition to Language*. Oxford, UK: Oxford University Press.

Steels, Luc, Martin Loetzsch (2008). Perspective alignment in spatial language. In Kenny Coventry, Thora Tenbrink, John Bateman (Eds.), *Spatial Language and Dialogue*. Oxford University Press.

Steels, Luc, Martin Loetzsch (2012). The grounded naming game. In Luc Steels (Ed.), *Experiments in Cultural Language Evolution*. Amsterdam: John Benjamins.

Tomasello, Michael (1995). Joint attention as social cognition. In Chris Moore, Philip J. Dunham (Eds.), *Joint Attention: Its Origins and Role in Development*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tomasello, Michael (1999). *The Cultural Origins of Human Cognition*. Harvard: Harvard University Press.

Tomasello, Michael (2001). Perceiving intentions and learning words in the second year of life. In Melissa Bowerman, Stephen C. Levinson (Eds.), *Language Acquisition and Conceptual Development*, 132–158. Cambridge: Cambridge University Press.

Van Looveren, Joris (1999). Multiple word naming games. In *Proceedings of the 11th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC '99)*. Maastricht, the Netherlands.

Vogt, Paul, Federico Divina (2007). Social symbol grounding and language evolution. *Interaction Studies*, 8(1), 31–52.

Wray, Alison (1998). Protolanguage as a holistic system for social interaction. *Language & Communication*, 18, 47–67.

Zadeh, Lotfi (1965). Fuzzy sets. *Information and Control*, 8, 338–353.