*Connection Science*
Vol. 20, No. 2-3, June-September 2008, 1–18

# Flexible Word Meaning in Embodied Agents

P. WELLENS†*, M. LOETZSCH‡ and L. STEELS†‡

†Artificial Intelligence Laboratory, Free University of Brussels, Brussels, Belgium
‡ Sony Computer Science Laboratory, Paris, France

(*v1.1 released January 2006*)

Learning the meanings of words requires coping with referential uncertainty – a learner hearing a novel word cannot be sure which aspects or properties of the referred object or event comprise the meaning of the word. Data from developmental psychology suggests that human learners grasp the important aspects of many novel words after only a few exposures, a phenomenon known as fast mapping. Traditionally, word learning is viewed as a mapping task, in which the learner has to map a set of forms onto a set of pre-existing concepts. We criticize this approach and argue instead for a flexible nature of the coupling between form and meanings as a solution to the problem of referential uncertainty. We implemented and tested the model in populations of humanoid robots that play situated language games about objects in their shared environment. Results show that the model can handle an exponential increase in uncertainty and allows scaling towards very large meaning spaces, while retaining the ability to grasp an operational meaning almost instantly for a great number of words. Additionally, the model captures some aspects of the flexibility of form-meaning associations found in human languages. Meanings of words can shift between being very specific (names) and general (e.g. "small"). We show that this specificity is biased not by the model itself but by the distribution of object properties in the world.

*Keywords:* word learning, lexicon acquisition, referential uncertainty, embodiment, language games, multi-agent simulations

## 1 Introduction

One of the greatest challenges in acquiring a lexicon is overcoming the inherent referential uncertainty upon hearing a novel word. This is because linguistic symbols embody a rich variety of perspectives – speakers use different words to draw the attention of the hearer to different aspects of the same object or event. Some of these contrasts are generality-specificity ("thing", "furniture", "chair", "desk chair"), perspective ("chase-flee", "buy-sell", "come-go", "borrow-lend") and function ("father", "lawyer", "man", "American") or ("coast", "shore", "beach") [Langacker, 1987]. Just from perceiving an object and hearing a word that supposedly describes that object, a word learner cannot know the intended meaning of the word. This problem is commonly related to the term "referential indeterminacy". Quine [1960] presented an example picturing an anthropologist studying the – unknown to him – language of a tribe. One of the natives utters the word "gavagai" after seeing a rabbit. How can, even after repeated uses of this word, the anthropologist ever come to know the meaning of "gavagai"? It could mean rabbit, an undetached rabbit part, food, running animal or even that it's going to rain. Children are very good at dealing with this problem. From the age of around eighteen months to the age of six years, they acquire on average nine new words a day (or almost one per waking hour). They can infer usable word meanings on the basis of just a few exposures, often without explicit training or feedback – a phenomenon that is known as fast mapping [Carey, 1978, Bloom, 2000].

Word learning is commonly viewed as a mapping task, in which a word learner has to map a set of forms onto a set of pre-established concepts [Bloom, 2000]. The implicit assumption is that learners have access to a number of potential meanings and need to choose (or guess) the correct one. Building on this assumption, several solutions to the problem of referential uncertainty have been theorized. One proposal is that the learner is endowed with several word learning constraints (or biases) that guide him towards the right mapping [see for example Gleitman, 1990, Markman, 1992]. Although the problem of referential uncertainty is acknowledged in this approach, it is also largely circumvented by claiming that learners

*Corresponding author. Email: pieter@arti.vub.ac.be

are able to almost instantaneously establish a mapping between a novel word and its meaning. Another suggestion proposes that learners enumerate all possible meanings the first time they are confronted with a novel word and prune this set in subsequent communicative interactions that involve the same word. This approach, while taking into account the problem of referential uncertainty, does not explain fast mapping. Smith et al. [2006] has shown that under the assumption of atomic word meanings, large vocabularies are learnable through cross-situational learning. But the time needed to grasp a usable meaning far exceeds the number of exposures as observed in children, especially when scaling to high dimensional meaning spaces. This is why often these two proposals go together: word learners use constraints to make a limited list of initial mappings and rule out all except one hypothesis later on.

Instead of characterizing a child as *identifying* the meaning of a word from a set of plausible possibilities, Bowerman and Choi [2001] envision the child as constructing and gradually shaping word meanings. 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]. Although in this view learners also make guesses at the meanings of novel words, they are different in nature. 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. "For example, many young children overextend words such as *dog* to cover all four-legged furry animals. One way they home in on the adult extension of this word is by hearing many four-legged furry animals called by other names such as *horse* and *cow*" [Tomasello, 2003, pp 73–74]. 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 apriori assumptions and even if they could, they still face a towering uncertainty in identifying the more subtle aspects of word meaning and use.

The problem of referential uncertainty differs depending on which of the above views on word learning is followed. In this article, we present a computational model for dealing with referential uncertainty that does not rely on enumerations of possible meanings or word learning constraints. Instead, we argue for truly flexible representations of meanings and mechanisms for shaping these word meanings through language use. We implemented the model in physical robotic agents that are able to perceive the world through their cameras and have mechanisms to engage in communicative interactions with other robots. Populations of these robots play *language games* [Wittgenstein, 1967, Steels, 2001] about objects in their shared environment (see Figure 2). These games are routinized interactions in which a speaker tries, using language, to draw the attention of a hearer to a particular object in a shared scene. The speaker and hearer give each other feedback as to whether this was successful and point to the intended object in cases of failure. This allows the population, over the course of many interaction, to self-organize a language for talking about physical objects. Note that agents are implemented such that they do not have access to internal representations of other agents – there is no meaning transfer, telepathy or central control.

A long history of experiments already exists on the emergence of communication systems in the language game paradigm, both in simulated worlds and with robots interacting in real environments. Since the early nineties, complexity has steadily increased in the agents' communicative task, and thus also in the nature of the coupling between form and meaning. One of the first models of lexicon formation was the *Naming Game* [Steels, 1995], in which simulated agents have to agree on names for pre-conceptualized individual objects. Technically, they had to establish one-to-one mappings between words and (given) symbolic representations without internal structure as illustrated in Figure 1a. The problem of referential uncertainty does not appear in the Naming Game – when a speaker points to an object, it is immediately clear for the hearer which individual concept to associate with a novel word. The main focus in the Naming Game was on the problem of how to reach lexical conventions and coherence in a population of interacting agents. Since each agent can invent new words, different words with the same meaning (synonyms) spread in the population, which poses a problem for reaching coherence. In a default naming game implementation agents keep different hypotheses about the meaning of a word in separate one-to-one mappings between names and individuals. Each mapping is scored and synonymy damping mechanisms, mainly based on

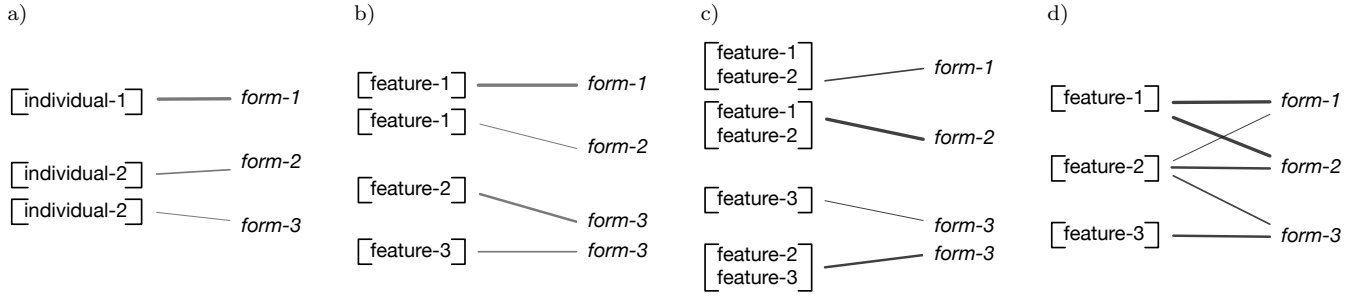



Figure 1. Increasing complexity in the nature of the coupling between form and meaning. Hypothetical example lexicons of one agent are shown for four different models of lexicon formation. Line widths denote different connection weights (scores). a) One-to-one mappings between names and individuals in the Naming Game. There can be competing mappings involving the same individual (synonyms). b) One-to-one mappings between words and single features in Guessing Games. Additionally to synonymy, there can be competing mappings involving the same words (homonymy). c) Many-to-one mappings between sets of features and words. In addition to synonymy and homonymy, words can be mapped to different competing sets of features that partially overlap each other. d) Associations as proposed in this article. Competition is not explicitly represented but words have flexible associations to different features that are shaped through language use.

lateral inhibition acting on these scores, were proposed to cope with the problem of incoherence.

When objects in the world are not represented as holistic symbols but instead different conceptualizations for the same object are possible, the problem of referential uncertainty appears. For example in *Guessing Games* such as the Talking Heads experiment [Steels and Kaplan, 1999], agents establish scored one-to-one mappings between words and perceptually grounded categories (or features, see Figure 1b). Hearers need to guess which sensory quality (size, color, position, etc.) a word is about and then choose an appropriate feature for that quality. In addition to synonymy, agents can adopt mappings to different features for the same word (homonymy). The amount of referential uncertainty, as measured by the number of different hypotheses, equals the number of different features of an object representation. One proposed mechanism to overcome this uncertainty is a word learning constraint: agents choose the sensory quality that is most salient in the scene (the difference between the topic and other objects in a scene is the highest for that quality). More prominently, cross situational learning [Siskind, 1996, Smith, 2005, De Beule et al., 2006, Smith et al., 2006, Vogt and Divina, 2007] has been shown to successfully solve the problem. In this approach, agents enumerate all possible meanings upon hearing a novel word and gradually refine this set by memorizing co-occurrences between forms and meanings. After many interactions, the mapping with the highest co-occurrence wins over the others and is used as the meaning of the word.

In natural language, words may refer to more than just single features such as [red] or [small]. One of the first models that allowed mappings between words and combinations of features as illustrated in Figure 1c was introduced by Van Looveren [1999]. It was shown to work when the number of object features is low. Since the meaning of a word can be any subset of the features of an object, referential uncertainty increases exponentially, as opposed to linearly in the guessing games outlined above. Suppose an object is represented by 60 features. The number of all possible subsets of these 60 features is $1.152921 \times 10^{18}$. Cross-situational approaches as outlined above become truly unfeasible since an agent cannot enumerate the long list of hypotheses which would be necessary to memorize co-occurrence relations. De Beule and K. Bergen [2006] have shown that when there is competition between specific (many features) and general (one feature) words, then general words will win over the specific ones because they are used more often – resulting again in an one-to-one mapping such as in Figure 1b.

In the model presented in this article, uncertainty is captured in the representation of word meaning itself (see Figure 1d). Instead of competing mappings that connect different sets of features to the same word, words have flexible connections to different features that are constantly shaped by language use. The model can be seen as an extension of cross situational learning, with the key difference that there is no enumeration of competing hypotheses and therefore the model can scale to very high dimensional hypothesis spaces. Learners *do not* take guesses, or choose from enumerations of possible meanings because the uncertainty is simply too great. The remainder of this article is structured as follows: in the next section we outline the experimental set-up that we use to test our approach. The model itself is explained in Section 3. Experimental results are presented in Section 4 and discussed in Section 5.

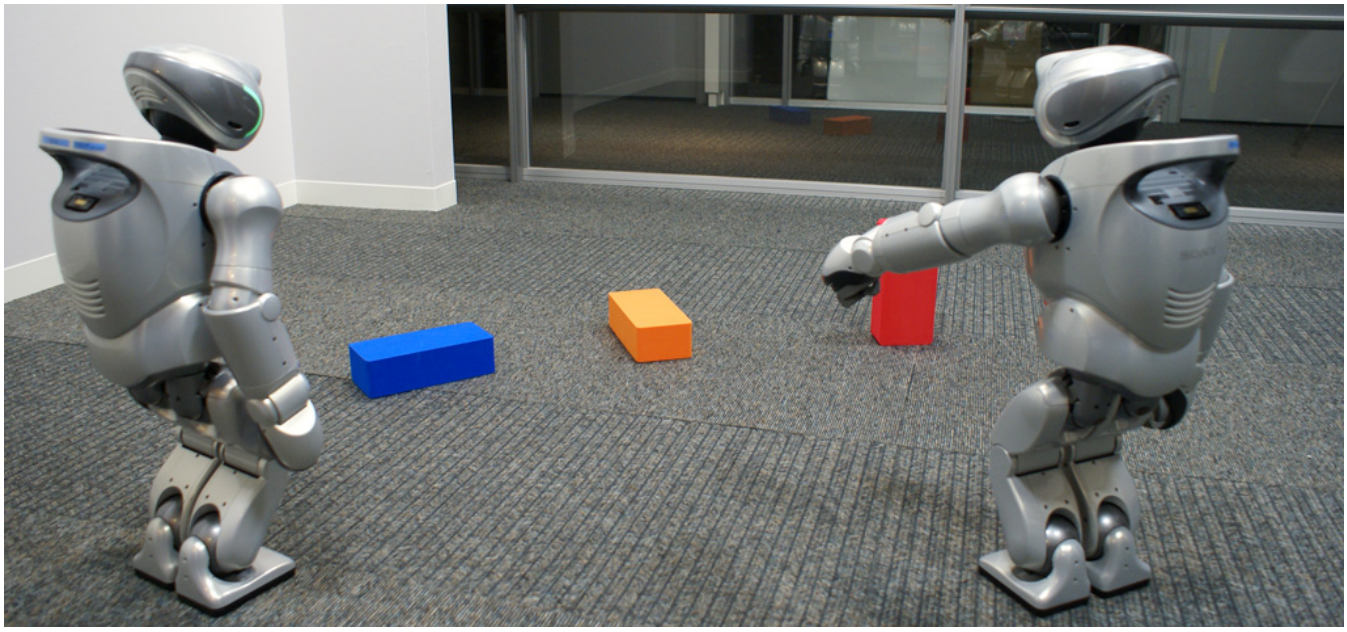
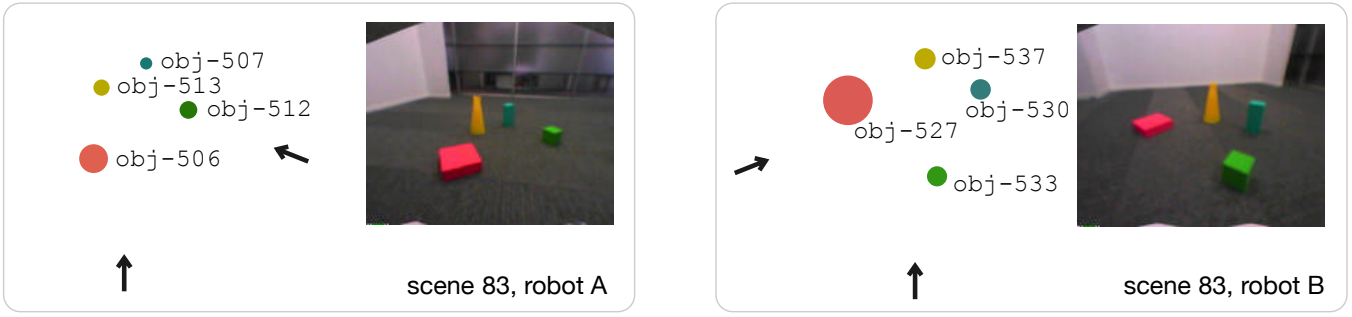


Figure 2. Sony QRIO humanoid robots play a language game about physical objects in a shared scene.

## 2 Interacting Autonomous Robots

The robotic set-up used in this experiment is similar to other experiments that investigate the cultural transmission of language in embodied agents [e.g. Steels and Kaplan, 1999, Steels and Loetzsch, 2008, see Steels, 2001 for an overview]. The experimental set-up requires at least two robots 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 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.

In this experiment, we use QRIO humanoid robots [Fujita et al., 2003] to test our model. The robots are about 60cm high and weigh 7.3 kg. They have a wide variety of sensors, including two cameras in the head, a microphone, and sensors in each motor joint to monitor posture and movement. Two QRIO robots are placed in an office environment that contains a set of geometric and toy-like colored objects (see Figure 2). Based on software developed for robotic soccer [Röfer et al., 2004], we developed a real-time visual object recognition system that is able to detect and track objects in image sequences captured by the built-in camera at the rate of 30 frames per second [Spranger, 2008]. The robots maintain continuous and persistent models about the surrounding objects using probabilistic modeling techniques. 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 3). From each such model, values on ten continuous *sensory channels* are extracted. In this experiment, these channels are the position of the object in an egocentric coordinate system (`x` and `y`), the estimated size (`width` and `height`), the average brightness (`luminance`), average color values on a green/red and a yellow/blue dimension (`green-red` and `yellow-blue`) and finally 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; `stdev-luminance`, `stdev-green-red` and `stdev-yellow-blue`). Note that the language model (see next section) does not depend on the choice of these 10 channels. Any other quality such as shape, texture, weight, sound, softness, etc. could be used, requiring techniques to construct it from the sensori-motor interaction with the environment. Channel values are scaled between 0 and 1. This interval 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 3 is `0.88`,




| obj-512 | obj-507 | obj-513 | obj-506 | obj-533 | obj-530 | obj-537 | obj-527 |
|---|---|---|---|---|---|---|---|
| *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* |
| 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 |
| luminance-2 | *luminance-2* | luminance-3 | luminance-2 | luminance-2 | *luminance-3* | luminance-3 | luminance-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 |
| yellow-blue-2 | yellow-blue-3 | yellow-blue-1 | *yellow-blue-2* | yellow-blue-2 | yellow-blue-3 | yellow-blue-1 | *yellow-blue-3* |
| stdev- | *stdev-* | stdev- | stdev- | stdev- | *stdev-* | stdev- | stdev- |
| luminance-3 | *luminance-4* | luminance-1 | luminance-3 | luminance-3 | *luminance-3* | luminance-1 | luminance-3 |
| *stdev-green-red-2* | stdev-green-red-4 | stdev-green-red-2 | stdev-green-red-4 | *stdev-green-red-3* | stdev-green-red-4 | stdev-green-red-2 | stdev-green-red-4 |
| *stdev-yellow-blue-3* | *stdev-yellow-blue-1* | stdev-yellow-blue-3 | stdev-yellow-blue-2 | *stdev-yellow-blue-4* | *stdev-yellow-blue-2* | stdev-yellow-blue-3 | stdev-yellow-blue-2 |

Figure 3. Visual perception of an example scene for robots 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, 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, differs. Those features that are different between the two robots are printed in italics.

so the assigned feature is `green-red-4`. We refer to the list of objects with their associated features as *sensory context*.

As mentioned earlier, populations of software agents play series of language games. All agents start with empty lexicons and have never before seen any of the physical objects in their environment. Since we have only two physical robots available and wish to model population sizes greater than two, they have to be shared. In each interaction two agents, randomly drawn from the population, embody the two robots to perceive their physical environment. At the start of the interaction, a human experimenter modifies the scene by adding/removing objects or by changing the position/orientation of objects. 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 draw the attention of the hearer to that object. For this he constructs an utterance, inventing new words when necessary and eventually uttering these words (these mechanisms are described in detail in the following section). The hearer interprets the utterance using his own lexicon and tries to find the object from his own perception of the scene that he believes to be most probable given the utterance. It could happen, however, that the hearer is confronted with a novel word or that his interpretation doesn't match any of the objects in his context. In this case, the hearer signals a communicative failure (by shaking his head). The speaker then points to the object he intended. When the hearer did understand the utterance, he points to the interpreted topic. The speaker then compares this object with the topic that he intended and either signals a communicative success (by nodding his head) or a communicative failure (by pointing to his intended topic). Finally, at the end of each interaction both agents modify their lexicons slightly based on the sensory context, the topic and the words used (*alignment*).

Since conducting thousands of such language games with real robots would be very time-consuming

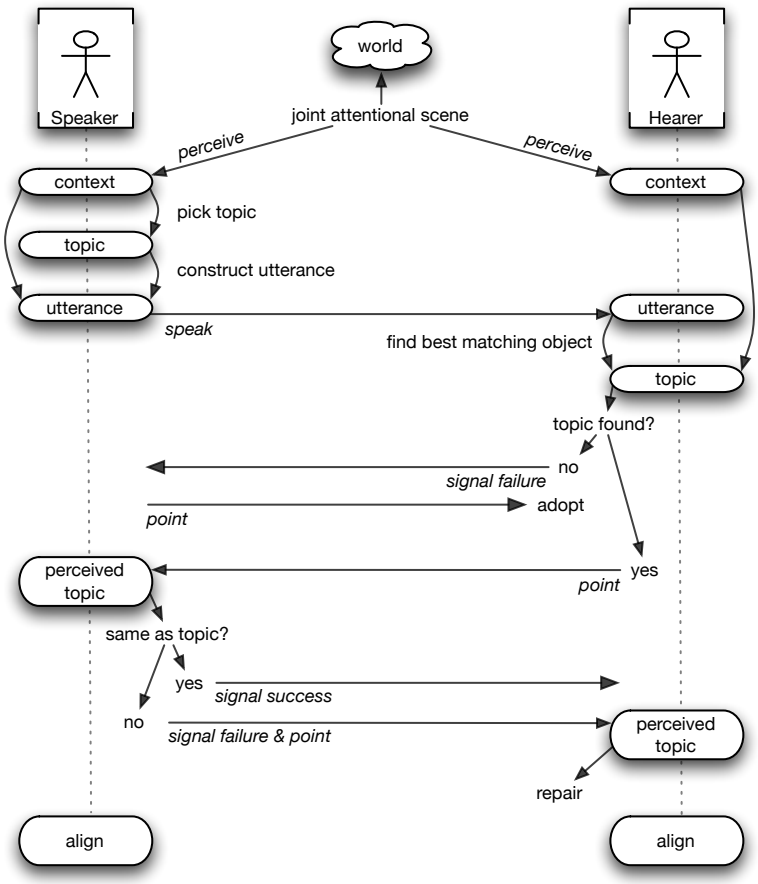



Figure 4. Flow of one language game. A speaker and a hearer follow a routinized script. The speaker tries to draw the attention of the hearer to a physical object in their shared environment. Both agents are able to monitor whether they reached communicative success and thus learn from the interaction by pointing to the topic of the conversation and giving non-linguistic feedback. Populations of agents gradually reach consensus about the meanings of words by taking turn being speaker and hearer in thousands of such games.

and also because we wanted repeatable and controlled experiments, we recorded the perceptions of the two robots (as in Figure 3) for 150 different scenes, each containing between 2 and 4 different objects of varying position and orientation out of a set of 10 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 A Flexible Model of Word Learning

As explained in the previous section, the vision system represents objects as sets of Boolean features. Though we are aware that such a representation lacks the richness needed to capture many interesting phenomena of human language and cognition, we believe this representation is sufficient for investigating the problem of referential uncertainty. Our language model itself is agnostic to the origins of the features. Using such a straightforward representation of objects and allowing the meaning of a word to be any subset of those features, the actual hypothesis space scales exponentially in the number of features. The first step towards a solution is to include uncertainty in the representation of word meaning itself. This is achieved by keeping an *(un)certainty score* for every feature in a form-meaning association instead of scoring the meaning as a whole. This representation is strongly related to both fuzzy set theory [Zadeh, 1965], with

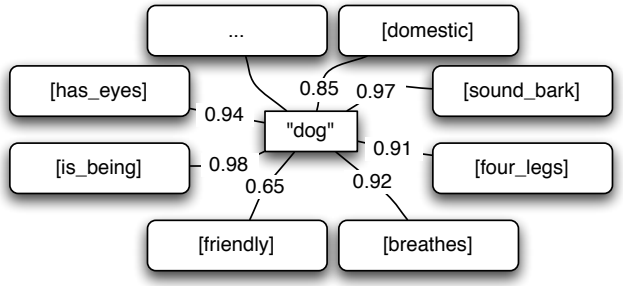



Figure 5. A possible representation for the word "dog" in English. Every feature associated with the form "dog" is scored separately.

the degree of membership interpreted as the degree of (un)certainty, and prototype theory [Rosch, 1973]. Although this representation is identical to a fuzzy set, in what follows, we refer to the representation as a *weighted set* to avoid confusion since we will redefine many set theoretic operations.

By allowing the certainty scores to change, the representation becomes adaptive and the need to explicitly enumerate competing hypotheses disappears. Thus, in contrast to most cross situational learning models it is not necessary to maintain and update a set of competing hypotheses. It follows that during production and interpretation (detailed in the following section) there is no need to choose between competing hypotheses since there is only a single hypothesis. As an example, the meaning of the word "dog" in Figure 5 *is* the *complete set* of scored associated features. Of course the features coming from the vision system in our experiment are much more basic then those depicted in Figure 5.

### 3.1 *Language Use in Production and Interpretation*

It is possible to define a weighted similarity measure for the above representation, taking the certainty scores as weights. Given two weighted sets of features as input, the measure returns a real number between $-1$ and 1, respectively denoting disjunction and equality. This weighted similarity measure lies at the core of the model and requires detailed elaboration but we first need to define some additional functions. Assume a function `Features(A)` that takes as input a weighted set $A$ and returns the normal set $B$ containing only the features from $A$, and another function `CertaintySum(A)` that takes as input a weighted set $A$ and returns a real number representing the sum of all the certainty scores. We can define the following operations as slight modifications from those of fuzzy set theory:

```
Function Intersection(A, B)
ForEach (feature & certainty) in A
    If Find feature in Features(B)
    then Add (feature & certainty) to intersection;
End ForEach;

Return intersection;
End Intersection

Function Difference(A, B)
ForEach (feature & certainty) in A
    If not Find feature in Features(B)
    then Add (feature & certainty) to difference
End ForEach;

Return difference;
End Difference
```

Note that function `Intersection` is *not* commutative in contrast to its definition in fuzzy set theory because it returns all shared features between $A$ and $B$ but takes the certainty scores from $A$. In what follows we will also use the union operation on fuzzy sets as defined in Zadeh [1965]. It takes the normal union of the two feature sets but when a feature appears in both $A$ and $B$ it takes the score with greater certainty.

Given these definitions we can define the weighted similarity measure as follows:

```
Function Similarity(A, B)
sharedSum ← CertaintySum(Intersection(A, B)) × CertaintySum(Intersection(B, A));
diffSum ← CertaintySum(Difference(A, B)) × CertaintySum(Difference(B, A));
similarity ← (sharedSum - diffSum) / CertaintySum(A) × CertaintySum(B);

Return similarity;
End Similarity
```

Given two weighted sets $A$ and $B$, `Similarity` first takes all shared features and all disjoint features from $A$ and $B$. By using the `CertaintySum` function we allow the certainty scores to weigh in. It is clear that sharing features is beneficial for the similarity while the opposite is true for features that are not shared. Intuitively, `Similarity(A,B)` will be high when $A$ and $B$ share many features with high certainty scores. Correspondingly, the result will be low when $A$ and $B$ have many disjoint features with high certainty scores. Some examples:

```
Similarity(((a 1.0) (b 0.5) (c 0.7)), ((a 0.5) (b 0.5) (c 0.7))) = (2.2 × 1.7 - 0 × 0) / 2.2 × 1.7 = 1
Similarity(((a 1) (b 1) (c 1)), ((d 1) (e 1) (f 1))) = (0 × 0 - 3 × 3) / 3 × 3 = -1
Similarity(((a 0.9)), ((a 1) (b 0.1) (c 0.2))) = (0.9 × 1 - 0 × 0.3) / 0.9 × 1.3 = 0.77
Similarity(((a 0.5) (b 0.5) (c 0.5)), ((a 0.5) (c 0.5) (d 0.5))) = (1 × 1 - 0.5 × 0.5) / 1.5 × 1.5 =0.33
```

We now have the ingredients to describe production and interpretation which both rely heavily on this similarity measure. As illustrated in Figure 4 the speaker, after picking a topic, must find an appropriate utterance to indicate the topic as clearly as possible. This process is called production and is implemented as follows:

```
Function Produce(context, topic, lexicon)
bestNewWord ← nil; // The current best new candidate word
utterance ← nil; // The utterance will gradually be constructed in here
productionScores ← nil;

Loop
   ForEach word in (lexicon \ words in utterance) do
      meaningOfUtterance ← FuzzySetUnion(ForEach word in utterance collect Meaning(word));
      meaningOfExtendedUtterance ← FuzzySetUnion(meaningOfUtterance + Meaning(word));
      objectSimilarities ← ForEach object in context
                              collect Similarity(meaningOfExtendedUtterance, object));
      topicSimilarity ← GetSimilarity(topic, objectSimilarities);
      closestOtherSimilarity ← Max(objectSimilarities \ topicSimilarity);
      Add (topicSimilarity − closestOtherSimilarity) to productionScores;
   End ForEach;
   bestNewWord ← word with highest score in productionScores;
   If ProductionScore(utterance with bestNewWord) > ProductionScore(utterance without bestNewWord)
   then Add bestNewWord to utterance;
   Else Break from Loop;
End Loop;

Return utterance;
End Produce
```

The `ForEach` loop will fill `productionScores` with a score for each unused word in the lexicon. This score represents the effect of adding the word to the current utterance by calculating its similarity to the topic and also taking into account its similarity to the rest of the context. For example if the topic is a red object, but all other objects in the context are also red it doesn't really help much to add the word "red". The `bestNewWord` is thus the word with the highest score in `productionScores`. If the productionScore for the utterance including `bestNewWord` improves upon that of the previous utterance it gets added to the `utterance`. If not, the search stops. In the end `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 context sensitive multi-word utterances and involves implicit, on-the-fly discrimination using the lexicon.

The most important effect of using a similarity measure is the great flexibility in word combination, especially in the beginning when the features have low certainty scores. Thanks to this flexibility the agents can 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).

Interpretation amounts to looking up the meaning of all the uttered words, taking the union of their (fuzzy) feature-sets and measuring the similarity between this set and every object in the context. The hearer then points to the object with highest similarity.

```
Function Interpret(utterance, context)
interpretedMeaning ← Union of all meanings for known words in utterance;
objectSimilarities ← ForEach object in context collect Similarity(interpretedMeaning, object);
topic ← object with highest score in objectSimilarities;
If similarityScore of topic > 0
then Return topic;
End Interpret
```

### 3.2 *Learning: Invention, Adoption and Alignment*

After finding the best possible combination of words to describe the topic, the speaker first tries to interpret his own utterance. In this process – which is called *re-entrance* [Steels, 2003] – the speaker places himself in the role of the hearer and thus can check for potential misinterpretations, allowing him to rephrase or remedy the utterance. When re-entrance leads the speaker to a different object than his own, which means that no combination of words can discriminate the topic in the current context, refinement of the lexicon is needed. 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. Because word meanings can shift, it might not be necessary to introduce a new word. Chances are that the lexicon just needs a bit more time to develop. Therefore high similarity between the meaning of the utterance and the topic translates to a lower likelihood of introducing a new word. In pseudocode the above process can be operationalised as follows:

```
Function Invention(utterance, topic, context)
interpretedMeaning ← Union of all meanings for known words in utterance;
interpretedTopic ← Interpret(utterance, context);
If interpretedTopic ≠ topic
then
   interpretedSimilarity ← Similarity(interpretedMeaning, interpretedTopic);
   topicSimilarity ← Similarity(interpretedMeaning,topic);
   randomNr ← Random(0 1); // A random number between 0 and 1
   If (interpretedSimilarity − topicSimilarity) > randomNr
   then
      newMeaning ← Features of (topic \ interpretedMeaning);
      newWord ← makeWord(randomString, newMeaning);
      Return newWord;
End Invention
```

When the hearer encounters one or more novel words in the utterance he needs a way to associate an initial representation of meaning with the novel forms. First the hearer interprets the words he knows 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 4), the hearer associates all unexpressed features with all the novel forms. Just as with invention the initial certainty scores start out very low, capturing the uncertainty of this initial representation. Excluding the features of the already known words is the only constraint shaping the initial representation. Note that there is no explicit enumeration of competing interpretations:

```
Function Adoption(utterance, topic, novelForms)
interpretedMeaning ← Union of all meanings for known words in utterance;
newMeaning ← Features of (topic \ interpretedMeaning)
ForEach form in novelForms do
   Add makeWord(form, newMeaning) to lexicon;
End Adoption
```

Flexible word use entails that in a usage event some parts of the meanings are beneficial (the shared ones) and others are not (the disjoint ones). If all features of the used meanings are beneficial in expressing the topic it would not be extension but instantiation, which is rather the exception than the rule. As Langacker

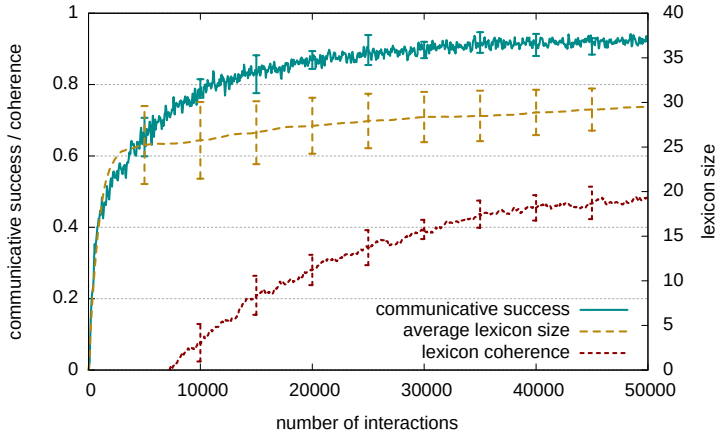



Figure 6. Dynamics of the language games in a population of 25 agents averaged over 10 runs of 50000 interactions. Values are plotted for each interaction along the x-axis. Communicative success: For each successful interaction (the hearer understands the utterance and is able to point to the object that was chosen as topic by the speaker), the value 1 is recorded, for each failure, 0. Values are averaged over the last 100 interactions. Average lexicon size: the number of words each agent knows is averaged over the 25 agents of the population. Lexicon coherence: This is a measure of how similar the lexicons of the agents are. For each word form known in the population, the similarity function described in section 3.1 is applied to all pairs of words known by different agents and the results are averaged. A value of 1 means that all 25 agents have identical lexicons, -1 means that they are completely different (each agent associates completely different feature sets to each word form) and a value of 0 means that the number of shared and non-shared features in the words of different agents is equal. Error bars are standard deviations across the 10 different experimental runs.

[2002] puts it, extension entails "strain" in the use of the linguistic items which in turn affects the meanings of these linguistic items. This is operationalised by slightly shifting the certainty scores every time a word is used in production or interpretation. The certainty score of the features that raised the similarity are incremented and the others are decremented. This resembles the psychological phenomena of entrenchment, and its counterpart semantic erosion (also refered to as semantic bleaching or desemantisation). Features with a certainty score equal or less than 0 are removed, resulting in a more general word meaning. In failed games 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:

```
Function Align(agent, topic, utterance)
topicFeatures ← Features(topic);
sharedFeatures ← Features(utterance) ∩ topicFeatures;
disjointFeatures ← Features(utterance) \ topicFeatures;

// Update association scores
ForEach word in utterance
   ForEach feature in Meaning(word)
      If feature in sharedFeatures
      then IncrementScore(word, feature);
      Else DecrementScore(word, feature); // Also removes features if score < 0
If not CommunicatedSuccessfully(agent)
then // Make words more specific, only the hearer does this
   ForEach word in utterance
      do Associate disjointFeatures to word;
End Align
```

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.

## 4 Experimental Results

We tested our model by letting populations of 25 agents play repeated series of 50000 language games. After only a few thousand games the agents reach their final lexicon size of 25 to 30 words (see Figure 6).

| form | agent 1 | | agent 2 | | agent 3 | | agent 4 | |
|---|---|---|---|---|---|---|---|---|
| *"murifo"* | x-4 | 0.46 | luminance-2 | 0.57 | luminance-2 | 0.38 | luminance-2 | 0.39 |
| | luminance-2 | 0.40 | yellow-blue-4 | 0.40 | x-4 | 0.29 | x-4 | 0.22 |
| | stdev-luminance-2 | 0.25 | green-red-2 | 0.40 | height-3 | 0.25 | yellow-blue-4 | 0.17 |
| | y-3 | 0.19 | x-4 | 0.32 | stdev-green-red-2 | 0.13 | green-red-2 | 0.17 |
| | stdev-yellow-blue-2 | 0.03 | height-4 | 0.19 | stdev-yellow-blue-3 | 0.08 | stdev-yellow-blue-2 | 0.10 |
| | | | stdev-green-red-2 | 0.12 | yellow-blue-4 | 0.08 | yellow-blue-2 | 0.10 |
| | | | stdev-yellow-blue-2 | 0.10 | y-3 | 0.08 | stdev-green-red-3 | 0.10 |
| | | | stdev-green-red-1 | 0.10 | | | stdev-luminance-4 | 0.10 |
| | | | stdev-luminance-2 | 0.10 | | | height-2 | 0.10 |
| | | | height-3 | 0.10 | | | y-3 | 0.10 |
| | | | width-4 | 0.10 | | | stdev-yellow-blue-3 | 0.06 |
| | | | y-3 | 0.10 | | | | |
| *"nuside"* | luminance-2 | 0.58 | yellow-blue-2 | 0.68 | yellow-blue-2 | 0.56 | yellow-blue-2 | 0.69 |
| | yellow-blue-2 | 0.49 | luminance-2 | 0.59 | stdev-yellow-blue-1 | 0.27 | luminance-2 | 0.36 |
| | stdev-yellow-blue-1 | 0.39 | width-2 | 0.31 | height-3 | 0.24 | stdev-yellow-blue-1 | 0.36 |
| | stdev-luminance-1 | 0.24 | stdev-yellow-blue-1 | 0.29 | | | height-3 | 0.24 |
| | height-3 | 0.19 | stdev-luminance-1 | 0.17 | | | stdev-luminance-1 | 0.15 |
| | stdev-green-red-1 | 0.17 | x-3 | 0.17 | | | width-2 | 0.13 |
| | y-3 | 0.17 | stdev-green-red-2 | 0.08 | | | y-3 | 0.13 |
| | x-4 | 0.17 | | | | | x-2 | 0.06 |
| | | | | | | | green-red-1 | 0.02 |
| *"migata"* | green-red-2 | 0.50 | luminance-2 | 0.40 | luminance-2 | 0.44 | luminance-2 | 0.49 |
| | luminance-2 | 0.48 | stdev-luminance-2 | 0.33 | x-4 | 0.38 | stdev-luminance-2 | 0.38 |
| | yellow-blue-4 | 0.39 | stdev-green-red-2 | 0.32 | stdev-luminance-2 | 0.21 | x-4 | 0.38 |
| | stdev-luminance-2 | 0.33 | stdev-yellow-blue-3 | 0.32 | yellow-blue-4 | 0.20 | yellow-blue-4 | 0.13 |
| | stdev-yellow-blue-3 | 0.30 | green-red-2 | 0.32 | green-red-2 | 0.10 | green-red-2 | 0.13 |
| | stdev-green-red-2 | 0.22 | x-4 | 0.32 | | | y-3 | 0.10 |

Figure 7. The meanings of the first three words of agent 1 (out of a population of 25 agents) and the corresponding meanings in the lexicons of agents 2, 3 and 4 after 10000 interactions. The numbers on the right side are scores of the association to the feature.

Also from very early on (at around interaction 10000), the agents communicate successfully in more than 80% of the cases. Please note that on average each of the 25 agents takes part in only 800 out of 10000 interactions and thus play only 4000 games in total. Although the agents can communicate successfully almost from the start, coherence is low (even negative) in the beginning, which means that the agents associate very different feature sets to each word form. Coherence continuously increases over the course of the following interactions and after 50000 games, communicative success has risen to 95%, indicating that the agents progressively align their word meanings.

To explain the very low initial lexicon coherence, Figure 7 lists the meanings of the first three words of agent 1 after 10000 interactions (communicative success $\approx$ 80%) and compares them with the meanings that agents 2, 3 and 4 connect to these forms. For each word, the features associated to it and the scores of the association are shown (sorted by score). It is immediately clear why lexicon coherence is so low in the population: each agent indeed associates drastically different feature sets of highly varying size to the same word forms. For example, all four agents associate different height information to the word "murifo": none for agent 1, `height-4` and `height-3` for agent 2, `height-3` for agent 3 and `height-2` for agent 4. The number of features connected to the word "nuside" ranges from three (agent 3) up to nine (agent 4). For nearly every word form, each agent associates at least one feature that no other agent connects to the same form. Words can even be associated to multiple features on the same sensory channel. For example, agent 4 has the features `yellow-blue-2` and `yellow-blue-4`, as well as `stdev-yellow-blue-2` and `stdev-yellow-blue-3` in its feature set for the word "murifo". The agents could not, however, communicate successfully if word meanings were not (at least) partially shared. Despite all the differences, the meanings of the three words in Figure 7 start to emerge: (almost) all agents associate `x-4`, `y-3`, `luminance-2` and `yellow-blue-4` to the word "murifo", giving it the meaning "far, left, uniformly dark and blue". For "nuside", the features `yellow-blue-2`, `luminance-2`, `height-3` and `stdev-luminance-1` are shared, meaning "high and uniformly yellow". The third word "migata" is associated by most of these 4 agents with `green-red-2`, `luminance-2`, `yellow-blue-4` and `x-4` ("far and turquoise"). This level of coherence is already enough for the agents to communicate successfully in many different contexts. Coherence continuously increases during the remaining 40000 interactions (see Figure 6), allowing the agents to communicate successfully in 95% of the cases after 50000 interactions.

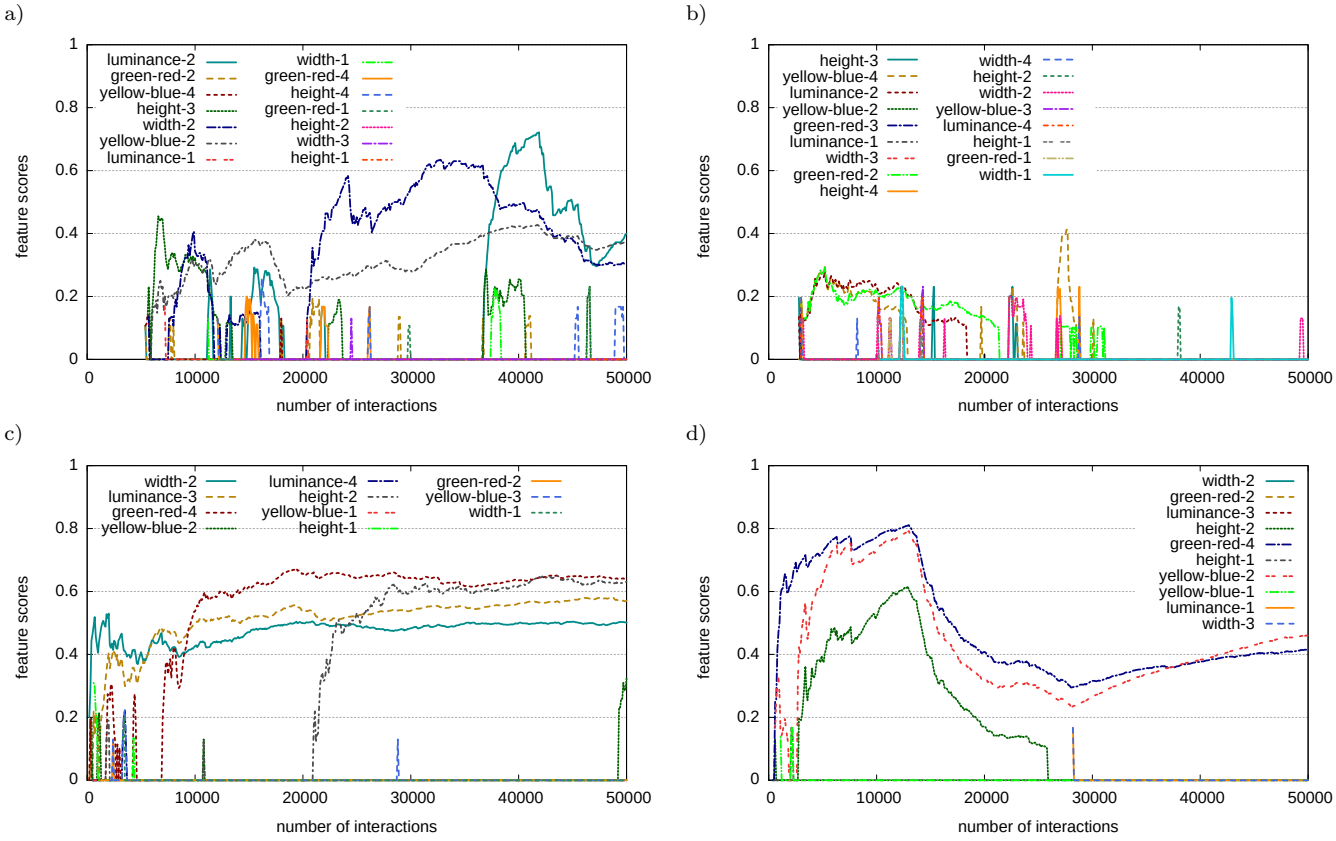



Figure 8. Examples of flexible word meanings. A population of 25 agents played 50000 language games. Each graph shows, for one particular word in the lexicon of agent 1, the strength of the association to different features. In order to keep the graphs readable, the agents agents have only access to a subset of the 10 sensory channels (`width`, `height`, `luminance`, `green-red`, `yellow-blue`).

In order to understand how the agents are able to align their initially very different lexicons, we looked at how the meanings of single words in one agent evolve over time. Not surprisingly, word meanings are extraordinary flexible and shift constantly. Figure 8 gives four examples of the changing association of word forms to different features. A word that constantly changes its dominant meaning is shown in Figure 8a. It is invented or adopted at around interaction 6000 and subsequently undergoes many meaning shifts. Over time, the highest association scores are to `height-3` (interaction 7000), `yellow-blue-2` (interaction 16000), `width-2` (21000 - 36000) and `luminance-2` (40000). Despite that, many other features become temporarily associated with the word, but are immediately discarded. The situation stabilizes towards the end, giving the word its final meaning "narrow, dark, yellow". In contrast, Figure 8b is an example of a rather unsuccessful word. The initial meanings disappear quite soon and at around interaction 5000, a stable set of 3 features arises. This meaning does not seem to spread over the population and the word loses all its features after 22000 interactions. Thereafter the agent does not use the word himself in production, but other agents in the population still use it, leading to new associations with features, which also ultimately remain unsuccessful.

In our model, words can be associated with any number of features. They can be very general, connected to only one feature (words such as "red" or "small"). They can also be very specific, similar to names, with connections to many features. And they can shift from general to specific and back. Despite some other associations that disappear very quickly, the word in Figure 8c 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 8d is an example of the opposite. It starts 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.

As mentioned earlier, human learners can infer usable meanings for a novel word after only a few

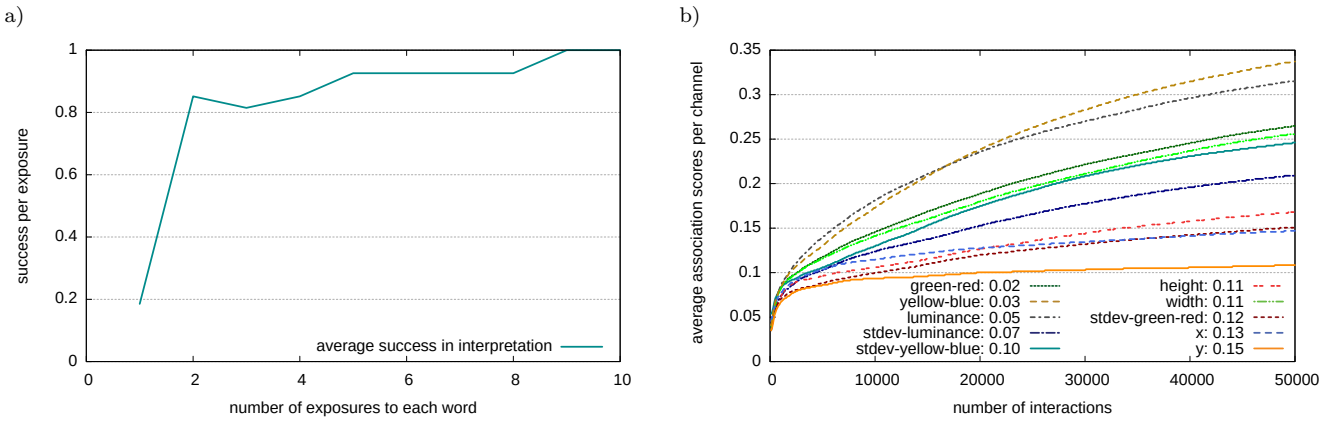



Figure 9. a) The interpretation performance of one new agent that is added to a stabilised population. For each word this agent adopts, the communicative success at the first, second, third etc. exposure is measured and averaged over all the words in the lexicon of that agent. b) The impact of the different perceptions on the lexicon: for each sensory channel, the average association score for channel features is shown, given all words in the population. In the legend, for each channel the average difference between the perception of robots A and B for all scenes in the data set are shown.

exposures. The graph in Figure 6 does not give us any insight on this issue, as it is about a population in the process of bootstrapping a lexicon. To investigate whether our model performs comparably to fast mapping, we added a new agent to a population that had already conventionalised a shared lexicon. The new agent only takes the role of a hearer, resembling a child born into a population that speaks a fairly stable language. The results, as depicted in Figure 9a, show that by the time of the second exposure 85% of the novel words lead to a successful interpretation. Further exposures gradually improve this result and by the tenth exposure all words result in a successful interpretation. This is even more surprising given that the other members of the population are unaware they are talking to a new agent, and thus use multi-word utterances, making it harder for the new agent to grasp the meanings of the words. In 20% of the cases, the new agent successfully interprets the utterance on the very first exposure to a new word because he understands enough of the other words to be able to point correctly.

When agents are embodied in physical robots, they have to deal with perceptual noise. The two robots view the scene from different angles and under different lighting conditions, leading to different perceptions of the same physical objects. However, the similarity in perception varies depending on the sensory channel. The average distance between the perception of a physical object between robots A and B on each sensory channel is shown in the legend of Figure 9b. This distance is computed by iterating over all objects of all scenes in the data set and for each sensory channel averaging the distances of the sensory values between the two robots. From the result we see that the most reliable sensory channels are `green-red` (average distance 0.02), `yellow-blue` (0.03) and `luminance` (0.05). The most varied channels show a very high level of difference, which makes them less suitable for successful communication: `y` (0.15), `x` (0.13) and `stdev-green-red` (0.12). The quality of a sensory channel is reflected in the agents' lexicons. Figure 9b shows the strength with which features are associated, for each sensory channel. This average score is computed for each channel by iterating over all the words in the population and averaging the scores of connections to features on that channel. The highest average scores are for features on the `yellow-blue`, `luminance` and `green-red` channels, the lowest for features on `y`, `x` and `stdev-green-red`. This corresponds perfectly to the average sensory differences on these channels, showing that the agents cope with perceptual differences by relying less on unreliable channels.

The world in which the robots interact "is structured because real-world attributes do not occur independently of each other. Creatures with feathers are more likely also to have wings than creatures with fur, and objects with the visual appearance of chairs are more likely to have funtional sit-on-ableness than objects with the appearance of cats. That is, combinations of attributes of real objects do not occur uniformly. Some pairs, triples, or ntuples are quite probable, appearing in combination sometimes with one, sometimes another attribute; others are rare; others logically cannot or empirically do not occur" [Rosch et al., 1976, p. 383]. For example, objects that are yellow also tend to be bright, tall objects are often also wide, and so on. This structure in the world is also reflected in the structure of the lexicons. Features that

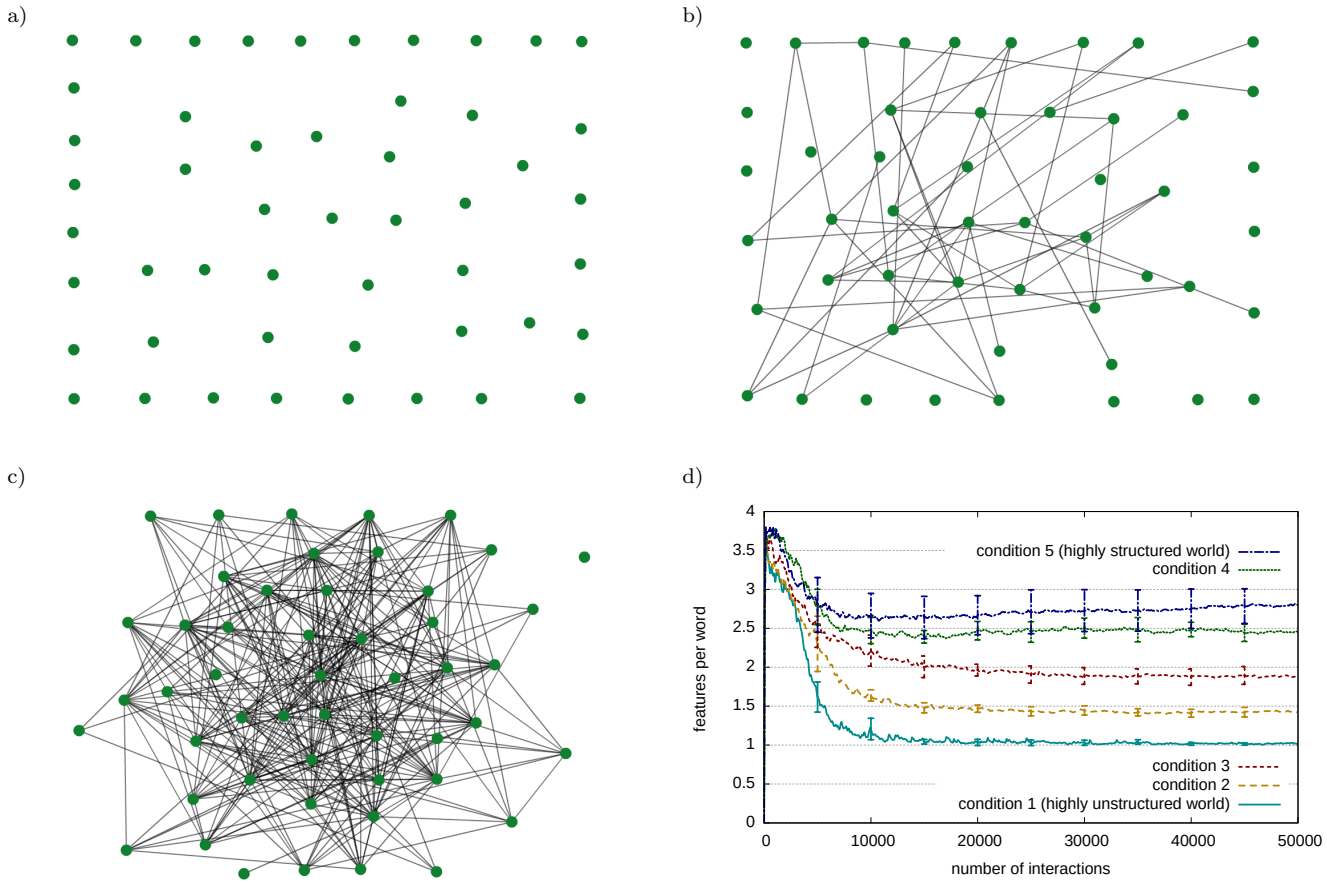



Figure 10. The effect of the amount of structure in a simulated world on the structure of the emerging language. Features are represented as nodes in a directed graph and feature nodes that are connected by edges will occur together in simulated perceptions of the world. a)-c) The co-occurrence graph used in condition 1 (highly unstructured world), condition 3 and condition 5 (highly structured world). d) The average number of features associated to each word for conditions 1 to 5. Values are averaged over all words in the population. Error bars are standard deviations over 10 repeated series of 50000 language games each.

co-occur often in will co-occur in the meanings of words. Since we cannot control the distributional properties of the object features in our previously recorded embodied data, we ran our model on a simulated world where such control is possible. We represented 50 features as nodes of a directed graph as shown in Figure 10a - 10c. Each node from index 1 to 50 was assigned a linearly decreasing probability for being attached to an edge. For different experimental conditions, a varying number of edges was added to the graph by connecting nodes randomly depending on their attachment probabilities. In each interaction, a sensory context similar to the ones in Figure 3 and consisting of five objects, each represented by 5 features is generated. Features were drawn by randomly selecting a node and taking also its neighbours having lower indices, until five features were chosen. The amount of structure in the world was controlled by the number of edges added to the graph: no edges in condition 1 (highly unstructured world, see Figure 10a), 20 in condition 2, 50 in condition 3 (see Figure 10b, 100 in condition 4, and 400 in condition 5 (highly structured world, see Figure 10c. We then ran series of 50000 language games with populations of 25 agents that are identical to the ones in the experiments above except that they use perceptions from the five differently structured simulated worlds, and compared the emerging lexicons as shown in Figure 10d. In condition 1, features co-occur completely randomly, making any attempt to capture re-occurring patterns in the world useless. This is clearly reflected in the lexicons of the agents. After about 10000 interactions, the words become essentially direct mappings of one feature to one form. On the other hand, objects in condition 5 show very high regularity, allowing the agents to create very specific words for specific objects. As a result, the average number of features covered in condition 5 is 2.75. The values for conditions 2–4 are between these extremes. This shows that in our model the specificity of words is not biased by the model itself but is a direct function of the structure in the world.

## 5 Discussion & Conclusion

In this article we introduced a new model of word learning for dealing with the problem of referential uncertainty. It does not rely on the simplifying assumptions made in previous models and instead builds on the idea that in order to tackle the uncertainty one must embrace it. We therefore argue for an adaptive representation of meaning that captures uncertainty at its core. This representation needs to be accompanied by a flexible manner of language use, captured in our model by defining a similarity measure. From this flexible use it follows that some parts of a meaning are beneficial, and others are not, opening the possibility for entrenchment and semantic erosion effects. Combining these ideas and repeating their effects over thousands of interactions results in a highly adaptive communication system with properties resembling some aspects also found in human languages. We tested the model in populations of physical robotic agents that engage in language games about objects in the real world. The results show that the model performs remarkably well, despite the difficulties arising from embodiment in robots and the high population size of 25 (compared to similar experiments).

In most previous experiments on lexicon formation, words are mapped to single components of meaning (individuals, categories or other features). Even in models where words can map onto sets of features, the dynamics are such that the agents finally arrive at one-to-one mappings between words and features [De Beule and K. Bergen, 2006]. This is due to the assumption that general words and specific words compete (against each other). Suppose there is a word "yellow" for the meaning [yellow], "fruit" for [fruit] and "banana" for [yellow, fruit]. When there are other objects in the world that are also [yellow] or [fruit], the words "yellow" and "fruit" will win the competition over "banana" because they are used more often. But in natural languages, different perspectives on the same object such as "yellow" and "banana" are clearly not competing but instead contribute richness. The model presented in this article does not have a bias toward one-to-one mappings between features and forms – words acquired by the agents can have any number of features associated to them. And pairs of words that share features such as "yellow" and "banana" do not compete because they are successfully used in different communicative situations. Finally, we showed that structure in the world and not the model itself, biases the structure of language and the specificity of words.

Although there is clear value in investigating the emergence of communication systems in simulated environments, we opted for an experimental set-up using situated robots. Presenting physical robots with a nontrivial communicative task in a rich and open-ended world prevented us from making unrealistic assumptions that were required in other models. For example, common scaffolding techniques such as direct meaning transfer between agents, or pre-conceptualized meanings, are not possible when robotic agents perceive real-world objects with initially unknown properties through cameras. Furthermore, not only the exponential uncertainty but also the complexity of our robotic set-up forced us to endow the agents with more flexible meaning representations and learning mechanisms. Both robots perceive the same scene from different angles so they can have drastically different views of the same object (for example the red block in Figure 3 has a much smaller width for robot A (`obj-506`) than for robot B (`obj-527`)). This makes guessing the meaning of a novel word even more difficult, because the intended meaning of the speaker might not even be among the different hypotheses constructed by the hearer. We argue that instead of trying to identify the meaning of a word by enumerating possible meanings, learners have to make an initial and necessarily uncertain representation that becomes refined over time. We showed that this actually happens in our model – different agents associate very different sets of features to the same words in the early interactions and then gradually reshape word meanings to reach coherence.

This "shaping of meanings" may make our model appear to be yet another variant of the cross-situational learning techniques as discussed above. But again we want to make very clear that there is a fundamental difference between structured (sets of features) and atomic (single feature) word meaning. We are not aware of any cross-situational learning model that allows meaning to be non-atomic or otherwise coping with exponential uncertainty. Smith et al. [2006] write: "Firstly, and most importantly, we have considered both words and meanings to be unstructured atomic entities" (p. 41). Furthermore, the agents in our language game experiments give each other non-linguistic corrective feedback, i.e. the speaker either confirms that the topic pointed at by the hearer was intended or points to the right topic. Lieven [1994] has shown that

children are able to learn many, and sometimes all, of their words without such social scaffolds. Vogt and Coumans [2003] have demonstrated that more natural "selfish games" which don't involve such a feedback are more difficult, albeit viable when tackled with cross-situational learning techniques. We did not test our model in such kind of interaction scenarios, but we speculate that the uncertainty stemming from missing feedback is of a lesser magnitude than the one resulting from exponentially scaling hypothesis spaces.

Finally, we want to clear some potential misunderstandings. First, we are not unsympathetic to the idea of word learning constraints, but we believe that constraints ony seem crucial when word learning is viewed as mapping. In this article we tried to show that by trading the mapping view for a more organic, flexible approach to word learning the constraints become less necessary. Second, some developmental psychologists emphasize human proficiency in interpreting the intentions of others [Tomasello, 2001] and our endowment with a theory of mind [Bloom, 2000] as main forces in word leaning. While being supportive of these ideas and even taking some for granted in our experimental set-up, it is important to understand that intention reading is not telepathy. These abilities might help in dealing with referential uncertainty, but they don't entirely solve the problem. Third, we do not take a position regarding the relation between the terms "word meaning" and "concept". Some researchers use these synonymously [Bloom, 2000], others advocate that they cannot be one and the same [Levinson, 2001]. In this experiment we did not investigate the subtle interplay between language, cognition and conceptual development but instead implemented a straightforward process from sensory experiences of objects to feature sets. This leads to the last point: since in our model agents have no other task but communicating, and therefore have no other internal representations beside word meanings, we cannot make any claims (pro or contra) regarding Whorf's thesis [Whorf and Carroll, 1956].

***Acknowledgments.*** 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 our experiments. We thank Michael Spranger for his indispensable help with the robotic set-up. 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 and the EU-FET ECAgents project (IST-2003 1940). Experiments are done in the Babel 2 framework, which can be freely downloaded from http://www.emergent-languages.org.

## References

P. Bloom. *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA, 2000.

M. Bowerman and S. Choi. Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. In M. Bowerman and S. C. Levinson, editors, *Language Acquisition and Conceptual Development*, pages 132–158. Cambridge University Press, Cambridge, 2001.

S. Carey. The child as word learner. In E. Wanner, M. Maratsos, M. Halle, J. Bresnan, and G. Miller, editors, *Linguistic Theory and Psychological Reality*, pages 264–293. MIT Press, Cambridge, Ma, 1978.

J. De Beule and B. K. Bergen. On the emergence of compositionality. In A. Cangelosi, A. Smith, and K. Smith, editors, *Proceedings of the 6th International Conference on the Evolution of Language*, pages 35–42, London, 2006. World Scientific Publishing.

J. De Beule, B. De Vylder, and T. Belpaeme. A cross-situational learning algorithm for damping homonymy in the guessing game. In L. M. Rocha, L. S. Yaeger, M. A. Bedau, D. Floreano, R. L. Goldstone, and A. Vespignani, editors, *Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*, pages 466–472, Cambridge, MA, 2006. MIT Press.

C. J. Fillmore. Scenes-and-frames semantics. In A. Zampolli, editor, *Linguistic structures processing*, pages 55–81. North Holland Publishing, Amsterdam, 1977.

M. Fujita, Y. Kuroki, T. Ishida, and T. T. Doi. 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)*, pages 960–967, Las Vegas, Nevada, October 2003.

L. Gleitman. The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55, 1990.

M. Haspelmath, M. Dryer, D. Gil, and B. Comrie, editors. *The World Atlas of Language Structures*. Oxford University Press, Oxford, 2005.

R. W. Langacker. A dynamic usage-based model. In M. Barlow and S. Kemmer, editors, *Usage-Based Models of Language*, pages 1–63. Chicago University Press, Chicago, 2002.

R. W. Langacker. *Foundations of Cognitive Grammar. Volume 1.* Stanford University Press, Stanford, 1987.

S. C. Levinson. Language and mind: Let's get the issues straight! In M. Bowerman and S. C. Levinson, editors, *Language Acquisition and Conceptual Development*, pages 25–46. Cambridge University Press, Cambridge, 2001.

E. Lieven. Crosslinguistic and crosscultural aspects of language addressed to children. In C. Gallaway and B. J. Richards, editors, *Input and Interaction in Language Acquisition*, pages 56–73. Cambridge University Press, Cambridge, 1994.

E. M. Markman. Constraints on word learning: Speculations about their nature, origins, and domain specificity. In M. Gunnar and M. Maratsos, editors, *Modularity and Constraints in Language and Cognition: The Minnesota Symposium on Child Psychology*, volume 25, pages 59–101, Hillsdale, NJ, 1992. Erlbaum.

W. Quine. *Word and Object*. MIT Press, Cambridge, MA, 1960.

E. Rosch. Natural categories. *Cognitive Psychology*, 7:573–605, 1973.

E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976.

T. Röfer, R. Brunn, I. Dahm, M. Hebbel, J. Hoffmann, M. Jüngel, T. Laue, M. Lötzsch, W. Nistico, and M. Spranger. GermanTeam 2004. In *RoboCup 2004: Robot Soccer World Cup VIII Preproceedings*, 2004. Extended version (299 pages) at http://www.germanteam.org/GT2004.pdf.

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

A. D. M. Smith. Establishing communication systems without explicit meaning transmission. In J. Kelemen and P. Sosík, editors, *Advances in Artificial Life Proceedings of the Sixth European Conference, ECAL 2001*, volume 2159 of *Lecture Notes in Artificial Intelligence*, pages 381–390, Berlin, 2001. Springer Verlag.

A. D. M. Smith. The inferential transmission of language. *Adaptive Behavior*, 13(4):311–324, 2005.

K. Smith, A. D. M. Smith, R. A. Blythe, and P. Vogt. Cross-situational learning: a mathematical approach. In P. Vogt, Y. Sugita, E. Tuci, and C. Nehaniv, editors, *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication*, pages 31–44. Springer Berlin/Heidelberg, 2006.

M. Spranger. World models for grounded language games. Diploma thesis. Humboldt-Universität zu Berlin, 2008.

L. Steels. Language games for autonomous robots. *IEEE Intelligent Systems*, 16(5):16–22, 2001.

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

L. Steels. A self-organizing spatial vocabulary. *Artificial Life*, 2(3):319–332, 1995.

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

L. Steels and F. Kaplan. Situated grounded word semantics. In T. Dean, editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99)*, pages 862–867, Stockholm, Sweden, August 1999. Morgan Kaufmann.

L. Steels and M. Loetzsch. Perspective alignment in spatial language. In K. R. Coventry, T. Tenbrink, and J. A. Bateman, editors, *Spatial Language and Dialogue*. Oxford University Press, 2008. To appear.

M. Tomasello. Perceiving intentions and learning words in the second year of life. In M. Bowerman and S. C. Levinson, editors, *Language Acquisition and Conceptual Development*, pages 132–158. Cambridge University Press, Cambridge, 2001.

M. Tomasello. *Constructing a Language. A Usage Based Theory of Language Acquisition*. Harvard University Press, London, UK, 2003.

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

M. Tomasello. *The Cultural Origins of Human Cognition*. Harvard University Press, Harvard, 1999.

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

P. Vogt and H. Coumans. Investigating social interaction strategies for bootstrapping lexicon development. *Journal of Artificial Societies and Social Simulation*, 6(1), 1 2003.

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

B. Whorf and J. Carroll. *Language, thought, and reality*. MIT Press, Cambridge, MA, 1956.

L. Wittgenstein. *Philosophische Untersuchungen*. Suhrkamp, Frankfurt, 1967.

L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.