Advances in Complex Systems
Vol. 15, Nos. 3 & 4 (2012) 1250031 (13 pages)

DOI: 10.1142/S0219525912500312

WORD AND CATEGORY LEARNING IN A CONTINUOUS SEMANTIC DOMAIN: COMPARING CROSS-SITUATIONAL AND INTERACTIVE LEARNING

TONY BELPAEME* and ANTHONY MORSE

*Centre for Robotics and Neural Systems,
University of Plymouth, A318 Portland Square,
Plymouth, PL4 8AA, United Kingdom
tony.belpaeme@plymouth.ac.uk

Received 24 August 2011
Revised 12 December 2011
Accepted 25 January 2012
Published 31 March 2012

The problem of how young learners acquire the meaning of words is fundamental to language development and cognition. A host of computational models exist which demonstrate various mechanisms in which words and their meanings can be transferred between a teacher and learner. However these models often assume that the learner can easily distinguish between the referents of words, and do not show if the learning mechanisms still function when there is perceptual ambiguity about the referent of a word. This paper presents two models that acquire meaning-word mappings in a continuous semantic space. The first model is a cross-situational learning model in which the learner induces word-meaning mappings through statistical learning from repeated exposures. The second model is a social model, in which the learner and teacher engage in a dyadic learning interaction to transfer word-meaning mappings. We show how cross-situational learning, despite there being no information to the learner as to the exact referent of a word during learning, still can learn successfully. However, social learning outperforms cross-situational strategies both in speed of acquisition and performance. The results suggest that cross-situational learning is efficient for situations where referential ambiguity is limited, but in more complex situations social learning is the more optimal strategy.

Keywords: Cross-situational learning; category acquisition; concept; language acquisition; cultural acquisition; social learning; language games.

1. Introduction

Language serves several functions all branching from or relating to its open-ended communicative function. It has for example been suggested that it plays a role in social cohesion (e.g., [11, 12]) and as a transmission mechanism for acquiring knowledge (e.g., [33]). As language externalizes internal knowledge, anyone interpreting a linguistic utterance has access to the internal state of the language user, albeit filtered through language. This process is central to cultural transfer of knowledge and to that of languages itself. One of the challenges in this respect has been to explain how language users acquire a language, a sub-problem being the acquisition of associations between words and their meanings. The problem of word-meaning acquisition was starkly defined by Quine [24], who argues that it is impossible from a theoretical point of view for a language learner to know what a novel word refers to, as there are an infinite number of possible referents for a word. As such, constraints need to be in place to restrict the number of potential referents for a novel word in order to solve the so called "indeterminacy of reference" problem. Obviously, children when acquiring a language seem to overcome this problem rather effectively, so the indeterminacy of reference is perhaps more of philosophical value. However, a complete picture of how children acquire the meaning of words is still lacking [7].

In developmental linguistics a range of possible mechanisms have been identified which language learners use to constrain the word-meaning acquisition process. Infants display a number of predispositions, beliefs and social strategies, such as the belief that object names are mutually exclusive [19] or that a novel name is associated with a nameless category [16]. Biases have been identified which help constrain the number of possible referents. The shape bias for example shows how infants have a preference for associating a novel word with shape rather than with other qualities [18]. Social strategies exist as well, such as joint attention [3, 32] and inference of the speaker's intent [1]. One strategy which could solve the indeterminacy problem and which has received considerable attention is *cross-situational learning* [2, 22]. In this, learners statistically pair words with meanings through repeated exposure to words and potential meanings: if a learner hears a novel word and perceives a number of potential meanings, then it can only assume that the word might be associated with any one of those meanings. However, if it hears the same word again, and perceives a different set of meanings, then it can assume that the word is associated with the intersection of the first set of meanings and the second set. If the learner keeps track of all meanings that co-occurred with a particular word, it is likely that there will be a moment where the referent of each word can be uniquely identified. Cross-situational learning does not require dyadic interaction: it is a passive learning strategy, whereby the learner inductively acquires the correct word-meaning mappings through repeated exposures to word and sets of meanings. The principle of cross-situational learning has been suggested as a potential acquisition mechanism [13, 22] and has been demonstrated in simulation [8, 15, 26, 28, 34, 35]. Smith and Yu [29] show how a cross-situational learning strategy is employed by 12 and 14-month old infants to learn the correct word-meaning mapping when repeatedly exposed to two meanings — one target and one distractor — and one word. Smith *et al.* [27] show how adults can employ cross-situational learning for complex exposures of up to 9 meanings and one word. All *in silico* and psychological studies however make one important assumption: that words and meanings are easily distinguishable. Either the meanings are discrete or the visual stimuli are sufficiently different as to assume that subjects cannot confuse

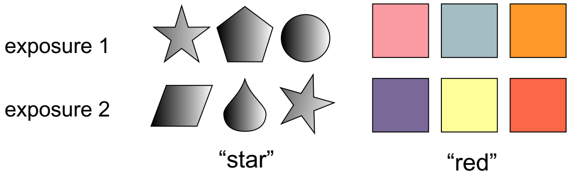


Fig. 1. Learning the meaning of a novel word is easier when the stimuli are perceptually distinct (left); when the stimuli have a graded membership, learning becomes more challenging (right).

the stimuli. In other words, there is no perceptual membership relation between the meanings and as such there can be no perceptual confusion between meanings — see Fig. 1 for an illustration.

In this study we wish to explore if cross-situational learning is still effective when the meaning space is continuous, i.e., perceptual stimuli have a membership relation to each other. Most natural categories [25] and perceptual categories — for example color, facial expressions — are continuous in nature and any two exemplars have a graded membership to each other. The challenge which cross-situational learning needs to solve is not only one of mapping a word to a meaning, but of distinguishing that meaning from possible distractors.

Cross-situational learning is a statistical learning method needing repeated exposures to learning experiences. As such it might be expected that it will not handle a situation where the meaning of a word is less clear. This study contrasts cross-situational learning with another, interactive, learning mechanism, in which the target meaning is specifically pointed out from the distractors. As such, the interactive learning is expected to perform better than cross-situational learning.[a]

The paper sets out to answer three questions:

(i) Is cross-situational learning still effective in a *continuous* semantic space?
(ii) What is the synchronic behavior of cross-situational learning compared to social acquisition?
(iii) What is the dynamic behavior (or diachronic) of cross-situational learning compared to social acquisition?

In Sec. 2 we first describe two implementations, one of a cross-situational learning model and one of a social learning model for *continuous* semantic domains. Next, in Sec. 3 we show quantitative results of how cross-situational and social learning cope with the task of word and meaning learning and simultaneously learning word-meaning associations. Finallly, in Sec. 4 we speculate on what these results might suggest for word-meaning acquisition in people.

[a] Cross-situational can be seen as an *unsupervised* learning method, as the learner does not receive feedback on its performance. Interactive learning, in contrast, is a *supervised* learning method, as the learner receives feedback on its inferences during each interaction with the teacher [14].

2. Simulation Models

This section describes the two models used in the study. The first model implements cross-situational learning for continuous semantic domains, as opposed to discrete semantic domains of earlier cross-situational learning models [8, 26, 28, 34]. The second model is of interactive, or social, learning for a continuous semantic domain, and is based on the language game methodology of Steels [30, 31]. Both models are multi-agent models, with $N \geq 2$ agents interacting, and rely on iterative interactions between the agents. Such models allow the study of diachronic evolution and the dynamics of various measures under different parameter settings. Models such as these have been used to study a variety of problems in language and cognition, such as the emergence of vowel systems [9, 21] or the acquisition of perceptual categories [4–6, 23].

The models share a common set-up. In a model N agents interact with each other (in the studies here $N = 2$). Each agent has a set of meanings M with individual meanings $m_i \in M \in [0, 1]^d$, with d being the dimensionality of the semantic space. An agent also has a set of words W, with individual words denoted as $w_j \in W$.

At each iteration of the simulation two agents are randomly selected to interact. They both observe a context C of $|C|$ stimuli selected randomly from $[0, 1]^d$. In the most general case, one agent acts as a speaker, the other as a hearer. In the implementation used here, one agent has a full set of meanings and words and will act as teacher; the other agent starts with an empty set of meanings and words and acts as learner.

We now describe two models, one model is used to study cross-situational learning (XSL) of meaning and words. The other model is used to study interactive learning (IL). The models have different methods of producing and interpreting linguistic labels: XSL does not rely on interaction (i.e., feedback) between agents: the learner passively absorbs learning experiences provided by the teacher. In contrast, in interactive learning there is a feedback loop between learner and teacher through which the learner received information on its performance.

2.1. *Cross-situational learning*

The cross-situational learning model is implemented using an instance-based approach, where the learner stores all instances of perceptual and linguistic input and then uses all the stored instances to produce and interpret words. When a context C is presented with $|C|$ stimuli $c_i, i = [1, |C|]$, the speaker produces a word w to refer to one of the stimuli. As the learner has no way of knowing which stimulus the speaker refers to, the learner stores all stimuli $c_i \in C$ and associates all with word w. If N contexts are presented, with $|C|$ stimuli in each context, then the learner stores a total of $N \times |C|$ stimuli and N words.

To *produce* a word for a stimulus, an agent uses a k-nearest neighbor (kNN) strategy, a simple nonparametric classification technique [20]. The agent finds the k nearest instances in its semantic space based on the Euclidean distance between the stimulus and each instance, and then uses the most frequent word among the kNNs.[b]

To *interpret* a word w produced for an unknown stimulus in a context containing $|C|$ stimuli, the hearer uses kNN to produce $|C|$ words for each stimulus in the context (as above). If one of the words produced by the hearer matches the word of the speaker, the associated stimulus is deemed to be the referent of w. If there is more than one match, then a stimulus is chosen at random from the matching stimuli.

2.2. *Interactive learning*

In interactive learning, we implement a variant of the *guessing game* [5]. In the guessing game the agents have a dyadic interaction when acquiring words and meanings. From a machine learning perspective, they engage in *supervised* learning. The algorithm does not use an instance-based approach as in individual learning: instead each agent stores a set of meanings M and a set of words W, words and meanings are associated with each other with association strength $a_{ij} \in A = C \times L \in [0, 1]$. Acquiring meanings, words and associations happens through an iterative process of repeatedly playing guessing games. In such a game, both agents perceive the same context C. The speaker chooses one stimulus at random, looks up a word for it and communicates this to the hearer. The hearer attempts to "guess" which stimulus the word refers to. If the guess is correct, the association between word and meaning is strengthened. If the guess is wrong, the association is weakened and the intended target is pointed out by the speaker — for further algorithmic details see [5].

To *produce* a word for a stimulus c_i in a context C, an agent finds the meaning $m_j \in M$ nearest to the stimulus based on the Euclidean distance D between the stimulus and categories: $\arg\min_j D(c_i, m_j)$. Next, it looks up the word w_k with the highest association $\arg\max_k a_{jk}$.

To *interpret* a word w, an agent finds the meaning m_i with the highest association with w. Finally it finds the stimulus c_j closest to m_i according to Euclidean distance: $\arg\min_j D(m_i, c_j)$ and points out the stimulus c_j to the speaker.

3. Results

The synchronic and diachronic performance of the models is described in the following section. Synchronic refers to the performance of the model *after* the learner has acquired a full set of meanings and words, and no further improvement in the learner's performance is observed.[c] We can study the model for a number of

[b] In k nearest neighbour, the algorithm is sensitive to the order k [17]. Relatively large values of k give good performance under noisy training data. When k is too large it can affect classification performance by taking into account instances of other classes. We empirically swept k from $[10, 20, \ldots, 100]$, which showed that kNN is not particularly sensitive in this context to the value of k. In our models we opted for a value of $k = 30$.

[c] "No further improvement" is implemented in the model as a change of 5% or less in performance measures over the last 100 interactions.

parameter settings, such as the complexity of the context and the number of meaning and word pairs that are being taught. Diachronic performance refers to the evolution of the learner's performance *during* learning.

We look at two performance metrics: *interpretation* performance and *production* performance. The interpretation performance measures how successful the learner is at correctly guessing the referent of a word of the teacher in a context. When calculating the interpretation performance of the learner, 100 contexts of $|C|$ stimuli are created, the teacher names a random stimulus in each context and the interpretation performance reports the ratio of correctly guessed stimuli by the learner. The production performance is calculated by letting both teacher and learner produce a word for a stimulus in 100 contexts of $|C|$ stimuli. The production performance is the ratio of identical named stimuli over the total number of named stimuli.

In the simulations described here, we use a special case of the models described in Sec. 2, with the number of agents $N = 2$. One agent will act as teacher and have a full set of meanings M_t and words W_t, with $|M_t| = |W_t|$ and each word uniquely associated with a meaning, i.e., the association matrix $A = I_{|M|}$. The second agent acts as learner, and has at the start of the simulation run an empty word and meaning set, and an empty association matrix, $M_l = W_l = A_l = \emptyset$. The dimensionality of the semantic space is set at $d = 2$: all stimuli and meanings are points in the two-dimensional $[0, 1] \times [0, 1]$ space. In the following section, all reported mean and standard deviations are calculated over 100 pseudorandomized simulation runs.

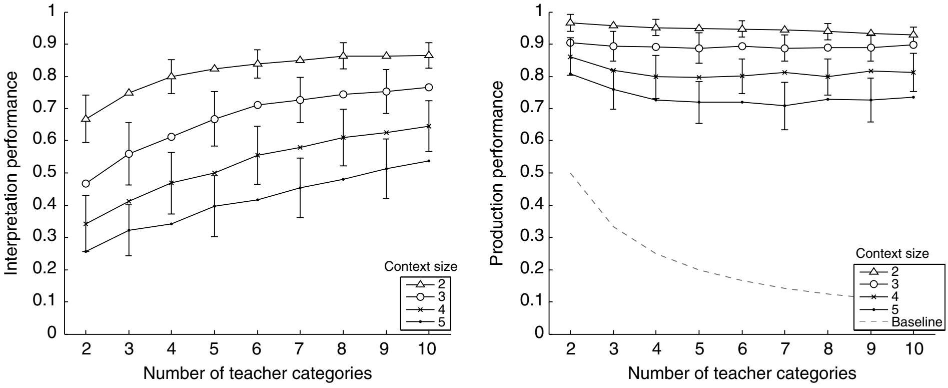
3.1. *Synchronic performance*

Figures 2 and 3 show the synchronic performance of respectively cross-situational learning and interactive learning, the size of the context $|C|$ is swept from 2 to 5 and the number of word-meaning pairs of the teacher is swept from 2 to 10. One-way ANOVA tests between different context sizes at $M_l = 10$ indicate that the differences between the means of the performance metrics are highly significant (all $p < 0.001$).

A number of observations can be made. First, in all instances, the performance of the learner is above chance. For the interpretation of words, chance-level performance is $1/|C|$, with $|C|$ being the context size. For production, in which both the teacher and learner produce a word for a stimulus and see if they agree, chance-level performance is given in Eq. (1), with $|L_t|$ and $|L_l|$ being the number of words of the teacher and learner.

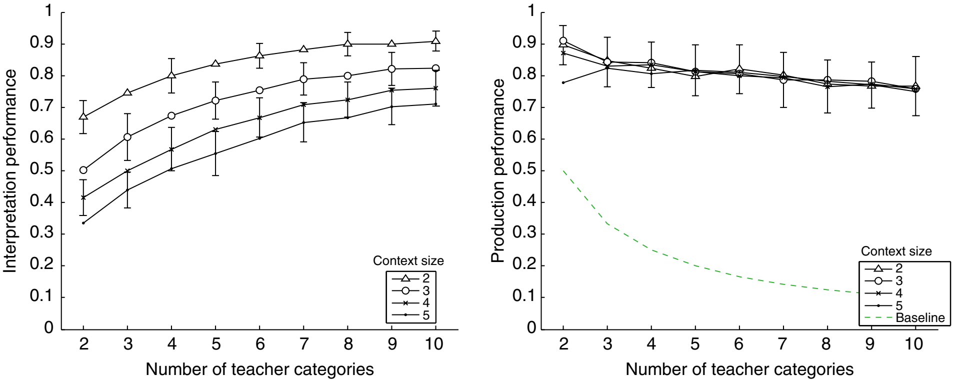
$$\min(|L_t|, |L_l|)/(|L_t||L_l|) \tag{1}$$

Second, the performance of the learner, both in interpreting and producing words, increases with the number of meaning-word pairs of the teacher and the size of the context. This seems counterintuitive, but can be understood when one



(a) (b)

Fig. 2. Interpretation (a) and production (b) performance of cross-situational learning for four different context sizes of $|C| = [2, 5]$. Error bars show standard deviation, the dashed line in (b) shows the chance-level production performance.



(a) (b)

Fig. 3. Interpretation (a) and production (b) performance of interactive learning for four different context sizes of $|C| = [2, 5]$. Error bars show standard deviation, the dashed line in (b) shows the chance-level production performance.

considers that a context is created randomly and might contain more than one exemplar categorized as the same meaning and thus being named with the same word. For a small context and small number of meanings of the teacher, it is more likely that the context will contain exemplars that will be named with the same word. This creates ambiguity which makes it hard for the learner to point out the correct target.

Finally, cross-situational learning seems to be more sensitive to the context size than interactive learning, see Figs. 2(b) and 3(b). As the context size increases, the performance drops. Indeed, more stimuli will make it harder for XSL to learn the

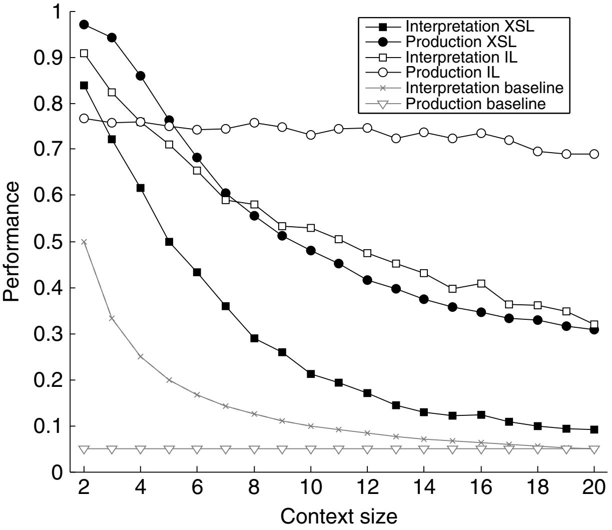
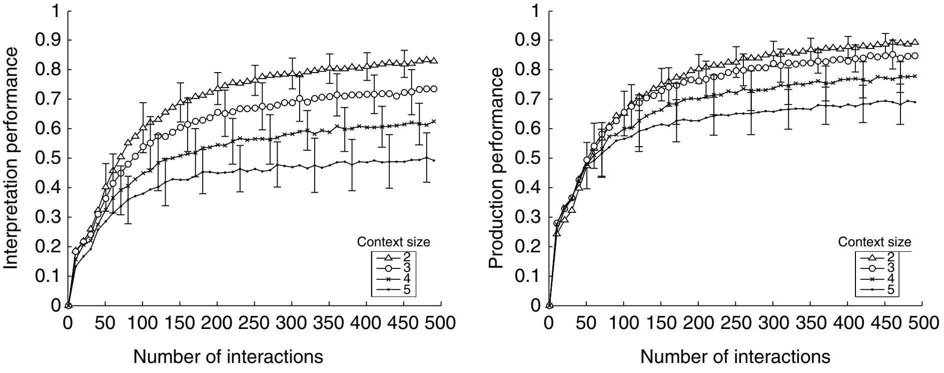


Fig. 4. Performance of cross-situational learning and interactive learning for various context sizes, with the number of meanings of the teacher kept constant at $|M_t| = 10$. The plot also shows the chance-level performance for both interpretation and production performance.

associations between words and meanings. This is further explored in Fig. 4 which shows the performance of both learning methods for different context sizes $|C| = [2, 20]$, with the number of word-meaning pairs for the teacher fixed at $|M_t| = |W_t| = 10$. Note that both learning methods perform above chance level, even for complex contexts of sizes of up to 20. The performance of cross-situational learning for small context sizes $|C| \leq 4$ matches or even outperforms that of interactive learning, but decreases fast with increasing context size. The performance of interactive learning is significantly less sensitive to increasing complexity of the context.

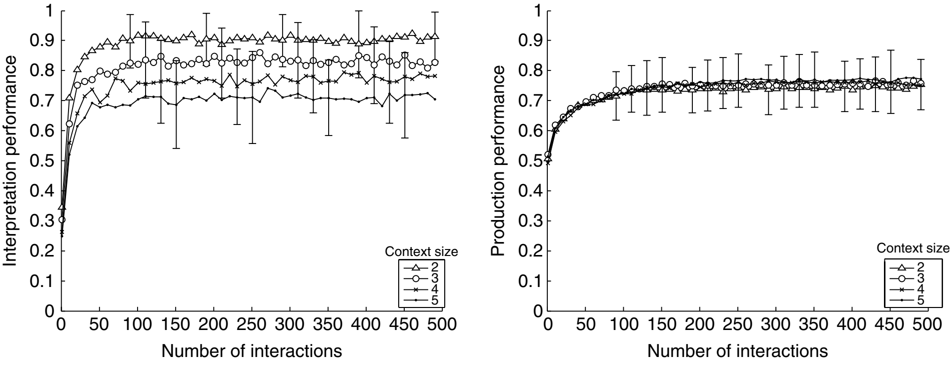
3.2. *Diachronic performance*

The diachronic characterization shows how a learner's performance evolves as it is exposed to an increasing number of learning experiences. At the start of a simulation run, the learner's performance — both for interpretation and production — will be at chance level. However, both for cross-situational learning and interactive learning, both performance metrics rise fast. Figure 5 show the interpretation and production performance of a learner employing cross-situational learning, again for different context sizes $|C| = [2, 5]$. The plots show the mean and standard deviation of 100 runs, whereby the learner is exposed to 500 learning exposures. The number of meanings (and associated words) of the teacher has been fixed at 10, all other parameters are identical as in Sec. 3.1. Both performance measures increase rapidly, but the increase slows down with increasing learning exposures. Increasing context sizes results in a decreasing performance (as per Fig. 4).



(a) (b)

Fig. 5. Diachronic interpretation (a) and production (b) performance of agents using cross-situational learning for different contex sizes $|C| = [2, 5]$.



(a) (b)

Fig. 6. Diachronic interpretation (a) and production (b) performance of agents using interactive learning for different contex sizes $|C| = [2, 5]$.

Figure 6 shows the diachronic performance for interactive learning for different context sizes $|C| = [2, 5]$. Again, both performance measures rise fast. As observed before, the production performance is not significantly sensitive to context size [Fig. 6(b)].

Figure 7 shows the difference between diachronic performance for cross-situational learning and interactive learning. Figure 7(a) shows the difference between the interpretation performance of the two learning methods, $IP_{IL} - IP_{XSL}$. The peak around interaction 50 shows how interactive learning initially performs faster and still outperforms cross-situational learning as the number of interactions increase. Figure 7(b) shows $PP_{IL} - PP_{XSL}$. Here interactive learning is faster than cross-situational learning, but for a small context $|C| \leq 3$ cross-situational learning after about 100 interactions outperforms interactive learning.

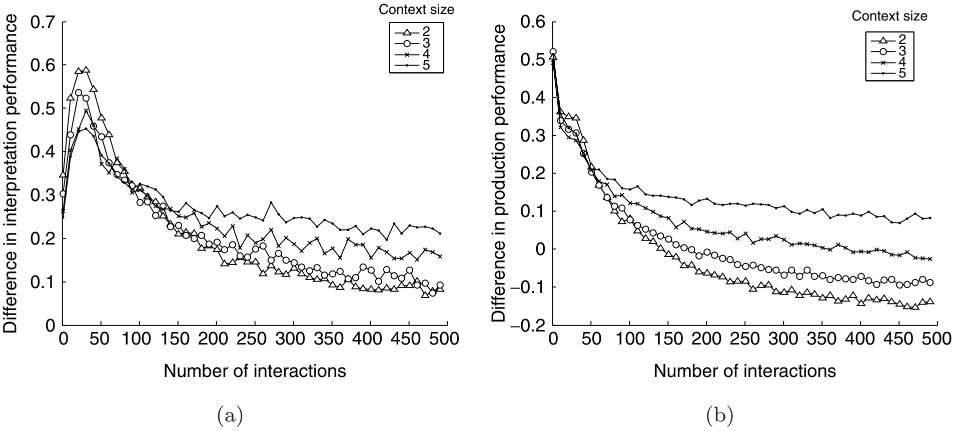


Fig. 7. Difference between diachronic interpretation (a) and production (b) performance of interactive learning (IL) cross-situational learning (XSL) and for different contex sizes $|C| = [2, 5]$. IL outperforms XSL for producing words, but XSL performs better for interpreting words in small contexts.

4. Discussion and Conclusion

Both models presented are simplifications of the processes of cross-situational learning and interactive learning, as such there is the possibility that the resulting behavior and the conclusions based on these are either too harsh or too forgiving. Interactive learning implements a feedback loop between the learner and the teacher, resulting in interactive learning being faster than alternatives where there is no feedback. If different methods for acquiring language and concepts would be pitted against each other, for example in evolutionary selection, then interactive learning mechanism would be favored when speed of acquisition is of importance. As such it is not unreasonable to suggest that interactive learning would be evolutionarily favored over other forms of non-interactive learning, such as cross-situational learning. However, cross-situational learning strategies can and are employed by infant and adult learners [27, 29]. Although cross-situational strategies are slower than dyadic strategies it has been observed that some acquisition processes, such as that of color categories, are also relatively slow; Vogt and Smith [34] suggested that cross-situational learning might be responsible.

The models used in this study do not use any biases or constraints to optimize learning. A number of constraints and biases exist [10], some of which have been identified in human learners, which help infants and adults disambiguate the referent of a word. For example, the novel name-nameless category (N3C) bias associates a novel label with the most unfamiliar stimulus [1] and could be used to speed up both learning approaches (see [27] for additional suggestions).

Note that a pure instance-based approach to individual learning, whereby all stimuli and words perceived by the learner are stored, is cognitively implausible: there is no neural correlate for KNN. As such, this study shows the ideal theoretical performance of a learner employing an instance-based learning strategy. Young children might employ a learning strategy whereby a number of contexts are stored and associated with words that is more parsimonious on memory, for alternatives see [8, 10].

Another unrealistic assumption in the model is the lack of noise. Noise would be introduced as soon as two or more teachers provide input to the learner, as it is unlikely that teachers have identical conceptual structures. The variation in their conceptual structures will result in noise on the input to the learner. Noisy input yet remains to be studied.

In this work we compared cross-situational learning and interactive learning for *continuous*, or graded membership, meanings. Fontanari and Cangelosi [14] compared both methods for discrete meanings. In their simulations, agents have to reach a consensus on how to name N object using H words. They show that while interactive learning outperforms cross-situational learning, the performance of both algorithms becomes identical in the asymptotical limit when both N and H go to infinity, but the ratio H/N remains finite. The performance of cross-situational learning and interactive learning for a large number of words and meanings in continuous semantic spaces remains to be studied.

Revisiting the questions set out in the introduction, we can confirm that cross-situational learning can deal with continuous semantic domains. However, the model suggests that performance for larger, and therefore more complex, contexts is significantly lower than for an interactive, or social, learning strategy. This comes as no surprise, as in interactive strategies referential ambiguity is greatly reduced by the teacher pointing out the referent of a novel word to the learner. This social strategy is used by young language learners and their parents as well, as exemplified by joint attention learning [32, 33]. However, as young and adult learners also employ cross-situational strategies, it seems plausible that humans use a variety of strategies for acquiring the meaning of words, with cross-situational learning strategies performing as well as social learning strategies when there is limited referential ambiguity.

Acknowledgments

This work is supported by the CONCEPT project (EPSRC EP/G008353/1) and the EU FP7 ITALK project. The author would like to thank Joachim De Greeff and the anonymous reviewers for their valuable feedback and discussions.

References

[1] Akhtar, N., Carpenter, M. and Tomasello, M., The role of discourse novelty in early word learning, *Child Dev.* **67** (1996) 635–645.

[2] Akhtar, N. and Montague, L., Early lexical acquisition: the Role of cross-situational learning, *First Lang.* **19** (1999) 347–358.

[3] Baldwin, D. A., Infant's ability to consult the speaker for clues to word reference, *J. Child Lang.* **2** (1993) 395–418.

[4] Baronchelli, A., Gong, T., Puglisi, A. and Loreto, V., Modeling the emergence of universality in color naming patterns, *Proc. Nat. Acad. Sci.* **107** (2010) 2403–2407.

[5] Belpaeme, T. and Bleys, J., Explaining universal colour categories through a constrained acquisition process, *Adap. Behav.* **13** (2005) 293–310.

[6] Belpaeme, T. and Bleys, J., The impact of statistical distributions of colours on colour category acquisition, *J. Cognit. Sci.* **10** (2009) 1–20.

[7] Bloom, P., *How Children Learn the Meanings of Words* (The MIT Press, Cambridge, MA, 2000).

[8] De Beule, J., De Vylder, B. and Belpaeme, T., A cross-situational learning algorithm for damping homonymy in the guessing game, in *Artificial Life X*, Rocha, L. M. *et al.* (eds.), (MIT Press, 2006), pp. 466–472.

[9] de Boer, B., *The Origins of Vowel Systems* (Oxford University Press, Oxford, UK, 2001).

[10] de Greeff, J., Delaunay, F. and Belpaeme, T., Human-robot interaction in concept acquisition: A computational model, in *IEEE International Conference on Development and Learning (ICDL 2009)*, Triesch, J. and Zhang, Z. (eds.) (IEEE, Shanghai, 2009), doi:10.1109/DEVLRN.2009.5175532.

[11] Dunbar, R., *Grooming, Gossip and the Evolution of Language* (Faber and Faber, London, 1996).

[12] Dunbar, R., The social brain: Mind, language and society in evolutionary perspective, *Annu. Rev. Anthropol.* **32** (2003) 163–181.

[13] Fisher, C., Hall, G., Rakowitz, S. and Gleitman, L., When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth, *Lingua* **92** (1994) 333–375.

[14] Fontanari, J. and Cangelosi, A., Cross-situational and supervised learning in the emergence of communication, *Interact. Stud.* **12** (2011) 119–133.

[15] Fontanari, J., Tikhanoff, V., Cangelosi, A., Ilin, R. and Perlovsky, L., Cross-situational learning of object-word mapping using neural modeling fields, *Neural Netw.* **22** (2009) 579–585.

[16] Golinkoff, R. M., Mervis, C. B. and Hirsh-Pasek, K., Early object labels: The case for a developmental lexical principles framework, *J. Child Lang.* **21** (1994) 125–155.

[17] Hall, P., Park, B. and Samworth, R., Choice of neighbor order in nearest-neighbor classification, *Ann. Stat.* **36** (2008) 2135–2152.

[18] Landau, B., Smith, L. and Jones, S., Object perception and object naming in early development, *Trends Cogn. Sci.* **2** (1998) 19–24.

[19] Markman, E. M., *Categorization and Naming in Children: Problems of Induction* (The MIT Press, Cambridge, MA, 1989).

[20] Mitchell, T., *Machine Learning* (McGraw-Hill, New York, 1997).

[21] Oudeyer, P.-Y., The self-organization of speech sounds, *J. Theor. Biol.* **233** (2005) 435–449.

[22] Pinker, S., *Language Learnability and Language Development* (Harvard University Press, Cambridge, MA, 1984).

[23] Puglisi, A., Baronchelli, A. and Loreto, V., Cultural route to the emergence of linguistic categories, *Proc. Nat. Acad. Sci.* **105** (2008) 7936–7940.

[24] Quine, W., *Word and Object* (The MIT Press, Cambridge, MA, 1960).

[25] Rosch, E., Natural categories, *Cognit. Psychol.* **4** (1973) 328–350.

[26] Siskind, J. M., A computational study of national techniques for learning word-to-meaning mappings, *Cognition* **61** (1996) 39–91.

[27] Smith, K., Smith, A. D. M. and Blythe, R. A., Cross-situational learning: An experimental study of word-learning mechanisms, *Cogn. Sci.* **35** (2011) 480–498.

[28] Smith, K., Smith, A. D. M., Blythe, R. A. and Vogt, P., Cross-situational learning: A mathematical approach, in *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication*, Vogt, P., Sugita, Y., Tuci, E. and Nehaniv, C. (eds.) (Springer Berlin/Heidelberg, 2006), pp. 31–44, doi:10.1007/11880172_3.

[29] Smith, L. and Yu, C., Infants rapidly learn word-referent mappings via cross-situational mappings, *Cognition* **106** (2008) 1558–1568.

[30] Steels, L., Evolving grounded communication for robots, *Trends Cogn. Sci.* **7** (2003) 308–312.

[31] Steels, L. and Belpaeme, T., Coordinating perceptually grounded categories through language. A case study for colour, *Behav. Brain Sci.* **24** (2005) 469–529.

[32] Tomasello, M., The role of joint attention in early language development, *Lang. Sci.* **11** (1988) 69–88.

[33] Tomasello, M., *The Cultural Origins of Human Cognition* (Harvard University Press, Cambridge, MA, 1999).

[34] Vogt, P. and Smith, A. D., Learning colour words is slow: A cross-situational learning account, *Behav. Brain Sci.* **24** (2005) 509–510.

[35] Yu, C. and Smith, L., Rapid word learning under uncertainty via cross-situational statistics, *Psychol. Sci.* **18** (2007) 414–420.