# Developing Lexicons in a Continuous Meaning Space

Avril Kenney

13 August 2010

**Abstract**

When people are acquiring language, it is not always clear from context which words refer to which objects, and the examples they see for any particular word will not all be the same. Previous approaches to the problem of learning or building a lexicon have characterized it as creating a mapping between a fixed set of words and a fixed set of meanings. This paper describes simulations of lexicon development using an agent-based model in which word-meanings are distributions over a continuous feature space; agents communicate about objects and infer the meanings of words based on the example objects they see. Agents in the model are able to develop a shared lexicon, and the meanings they reach are efficient given what objects tend to be present. Even when the objects are entirely random, the lexicons that develop tend to result in a near-optimal probability of communicative success. If new agents are being introduced into the population, the lexicon will continue to change, and will have periods of stability punctuated by periods of fluctuation.

# Introduction

## Children learning words

A child acquiring language, even relatively straightforward aspects like names for objects, for the most part does not learn it by being shown single objects paired with their labels. Children learn merely by being exposed to speech in the presence of real-world situations. In some cases there are cues to reference, such as pointing or eye gaze, but in many cases the nonlinguistic context does not disambiguate what is being referred to.

One mechanism for dealing with the problem of referential uncertainty is cross-situational learning. Although it might not be possible to infer the meaning of a word from a single context, a learner can use information from multiple contexts, looking at what was present across all of the instances that he has heard a particular word. It has been shown that even when no single occurrence of a word makes its meaning unambiguous, children are able to learn word–referent mappings cross-situationally (Smith & Yu, 2008).

Another important fact of word learning is that most words do not refer to exactly the same thing each time they are used. Mapping words to meanings is therefore more complicated than mapping names to individual objects: in addition to figuring out which word goes with which object in a particular context, the learner must figure out what objects a word is appropriate for in general. Tenenbaum and Xu (2000) show how people can infer the correct level of generalization for the meaning of a word based on seeing only a few examples.

## Efficiency of language

Despite all their apparent ambiguities and imperfections, the lexicons of human languages are fairly efficient for communication.

Languages often contain words that are supposedly synonyms, different words that have the same meaning, but it is quite uncommon for two words to have *exactly* the same meaning. This makes sense in terms of efficiency, because in different contexts it could be appropriate to use words with similar but slightly-different meanings, whereas the presence of two words that can be used completely interchangeably does not contribute any communicative value.

Lexical ambiguity can arise from polysemous words, for which there are multiple possible meanings of the same phonological wordform (for example, “type”: I can *type* words on the keyboard or talk about different *types* of words). But although such words when encountered individually would be ambiguous, they are almost always disambiguated by context. A corpus analysis by Gale, Church, and Yarowsky (1992) showed that when a polysemous word appears more than once within a discourse, its occurrences have the same sense in 98 percent of cases. A similar analysis found that about 95 percent of the time, words with two meanings can be disambiguated by linguistic context such as adjacent words or syntactic relationships (Yarowsky, 1993). By reusing the same phonetic forms for multiple meanings, languages reduce the number of words that have to be remembered, and because the meanings occur in dissimilar contexts, there is little ambiguity created.

Even for meanings that are not clearly segmented by the nonlinguistic world, natural languages' allocation of words to meanings is efficient. One well-studied example of a continuous space that is divided into different lexical items by different languages is color. A study of color naming in 110 languages found that when the perceptual features of color space are taken into account, a language's set of color terms tends to maximize within-item similarity and minimize between-item similarity (Regier, Kay, & Khetarpal, 2007). The same pattern was also found in the terminology for spatial relationships in 9 languages (Khetarpal, Majid, & Regier, 2009). These results indicate that even when there are no inherent boundaries in a space of meanings, natural languages generally partition the space in non-arbitrary, useful ways.

## Modeling lexicons

Previous modeling studies have shown that it is possible for agents communicating with each other to create a shared lexicon that has some of the important properties of human language.

Many such models take the form of agents playing "language games." For example, in one experiment (Steels, 1996), pairs of agents take turns communicating about a known referent with a known set of distinctive features (that differentiate it from other possible referents), updating their lexicons to associate each word they hear with the distinctive features of the referent, and creating words when none are appropriate. The agents eventually reach a point at which all communications are understood. Because words are created to refer to distinctive features, only the relevant meanings are lexicalized, and a single word can encompass multiple meanings.

Nowak and Krakauer (1999) show that in order for a shared lexicon to develop, either communication must be beneficial to both participants, or they must cooperate with each other. In their model, agents learn to pair up a set of sounds with a set of meanings, and successful communications lead to higher fitness, allowing agents to live longer and reproduce more. If some meanings are more valuable to communicate than others, then the individuals who have words for these meanings will have a higher fitness, so if the number of words is limited, the resulting lexicon will use words for the most valuable meanings.

Nowak, Plotkin, and Krakauer (1999) describe a mathematical model in which agents learn a lexicon by building an "association matrix," where each entry is the number of times a given word has been associated with a given meaning. This model provides a reason for why a single word can have multiple meanings but it is uncommon for multiple words to have the same meaning: a polysemous word will consistently be used to refer to objects with all of its meanings, so for each meaning a new agent will learn consistently the word for it; but for an object with two possible words, sampling instances will generally cause one of the words to be used more frequently by chance in the examples seen by a learner, so the learner will use one of the words more often for the meaning, and by repeating this process the lexicon will reach a stable state in which only one of the words is used.

If agents are learning mappings and there are no errors in learning, an unchanging population will eventually converge on a stable lexicon, without necessarily having reached an optimal state (for example, a single word could be used for two separate meanings, decreasing its ability to convey information). But when there is some probability of error, lexicons are able to keep changing and will not remain permanently in a single state. If the population is also changing, the lexicon will exhibit punctuated equilibria, with periods of competition among periods of stability (Steels & Kaplan, 1998).

Vogt and Divina (2007) combine multiple learning mechanisms into a complex model of "social symbol grounding" that includes joint attention, feedback, cross-situational learning, and contrast (the fact that a novel word is unlikely to have the same meaning as a known word). The model differs from others in that it contains more of the real-world factors affecting language acquisition. They observe that joint attention and cross-situational learning are important for enabling word learning, whereas feedback and contrast matter less. This makes intuitive sense because the more important factors are those that enable learners to figure out which object a speaker is talking about.

## Inferring meanings

The problem of learning meanings from real-world examples has been explored in the study of "physical symbol grounding," in which symbols such as words are to be associated with patterns in sensory input such as unprocessed video from a camera.

Roy (1999) created a system that uses raw data from a microphone and a camera to learn a lexicon. It recognizes regularities in speech and in visual input, and forms associations based on the mutual information between the two modalities. Tested on recordings of real-life parent–child interactions, it was able to correctly learn words for simple objects.

Another system for learning meanings from sensory input is described by Steels and Kaplan (2000), who use mobile autonomous robots that interact with human teachers. The robots use object-classification techniques and verbal input from the teachers to learn to recognize objects associated with speech sounds. Their results emphasize the importance of social learning: the behavior of a teacher helps to direct the attention of the robots to what is being talked about, and constraints on what classifications can be learned are provided by social experience rather than just visual properties.

### Differences from the present work

The present work differs from previous models of lexicon creation in that the objects being communicated about are drawn from a continuous space rather than a finite set. Learning the lexicon therefore requires generalizing from examples rather than forming mappings to known meanings. In addition it includes cross-situational learning, enabling words to be learned even when there are multiple objects in a context and the referent is not specified.

It also differs from systems that learn words while using sensory input to uncover the structure of the world, because in this model the meanings are derived and learned primarily from linguistic input.

## Model

Simulations were performed using an agent-based model written in NetLogo. Agents represent people, each of whom has a lexicon, which is a mapping from symbols (wordforms) to probabilistic distributions over the space of possible meanings. (The representation of the wordforms themselves is not relevant in this model; they are internally represented as integers.) The agents communicate about "objects," which are represented as points in a continuous two-dimensional feature space, or "meaning space." An agent's representation of the "meaning" of a word is a set of Gaussian distributions (one for each dimension of the space).

Communications take place between pairs of agents. At each timestep, each agent communicates to each other agent with a certain probability (for the simulations reported here, this probability was constant (.1) that was the same for all pairs of agents). A set of one or more objects (points in meaning space) is generated, and both the speaker and the listener can "see" them (know the exact values of their features). The objects can be generated from a set of "real" meanings (in which case no two objects in a context will come from the same meaning), or they can be generated uniform-randomly from the space (Figure 1). (In all results described here, the number of objects in each context is randomly selected with an equal probability of being 1, 2, or 3.) The speaker agent randomly selects one to refer to, then chooses a word to say in order to get the listener agent to choose the correct object of the ones present. Which word he chooses is based on how well the word describes the referent (the probability of the referent given the word), how well the word describes the other objects present, and the proportion of previous communications using this word that have been successful. If none of the words in the speaker's lexicon is appropriate (i.e., the object is more than 2.5 standard deviations from the mean in at least one dimension, or the best word for the object has usually not been understood correctly in the past), with a certain probability he will make up a new word. The new word is centered on the referent object, and a default standard deviation is used for all dimensions. The speaker receives feedback so that he knows whether the listener chose the object he intended to refer to.

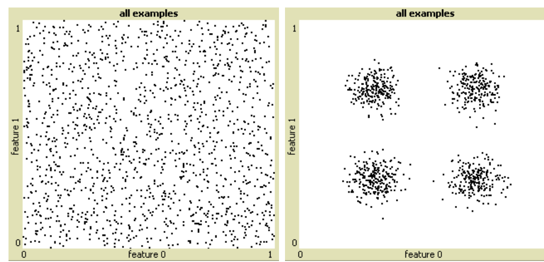



Figure 1: Points representing example referents in the space of meanings. Left: objects are generated uniform-randomly. Right: objects are generated from four "real" meanings.

When a listener hears a word, he selects which of the objects present he thinks the word refers to. If it is a word he has in his lexicon, he computes the maximum (over dimensions) number of standard deviations that each object is from the mean of the word, and chooses the object with the minimum value of that quantity. Otherwise, if he does not know the word, he guesses an object at random. Each agent keeps a list, for each of the words in his lexicon, of the example objects that he has heard that word used to refer to (agents have a limited memory, remembering only a fixed number most recent examples; in the simulations here, memory was 100). Every time he sees a new example for a word, he updates his meaning of that word: an agent's lexicon entry for a word has the means and standard deviations computed from the examples in his example list.

In the *ostensive learning* condition, the listener is given feedback as to which object was the intended referent, so he simply updates his example list with the known referent.

In the *cross-situational learning* condition, agents are not given direct feedback; they keep track of all the objects that have been present in each of the contexts in which they have heard each word, and infer which objects a word refers to based on the overlap between the different contexts. Because none of the contexts will contain objects with exactly the same features, the overlap cannot be computed by looking for individual points that occur in multiple contexts. Instead, whenever an agent hears a word, he selects the object closest to that word, and computes a score based on how probable the object is given the what he thinks the word means and how many times he has heard the word before. If this score is past a certain cutoff, he will reanalyze to find a more accurate meaning for the word, by looking at each object in the present context and using it to form a cluster consisting of one object from each of the contexts in which he has heard the word, and then choosing the tightest cluster.

A limit can be placed on the number of words that each agent is allowed to have in his lexicon. For a maximum lexicon size $N$, an agent keeps in his lexicon only the $N$ most recent words he has heard (or made up).

At each timestep, there is a fixed probability that an agent will die and be replaced by a new agent, who does not yet know any words. This process is independent of the communicating, and all agents have the same probability of being replaced. There is no "fitness" – an agent's lifespan is unrelated to his ability to communicate – and there are no stratified generations.

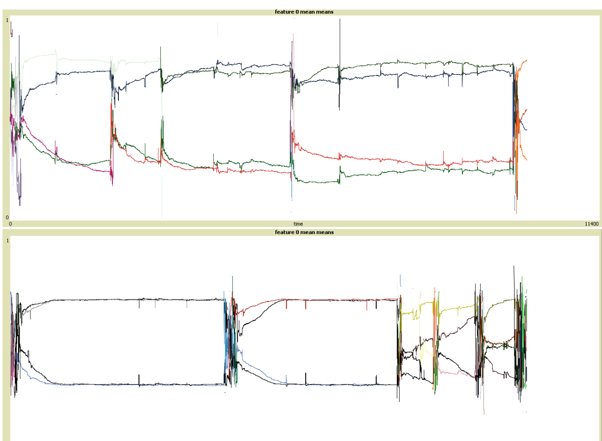
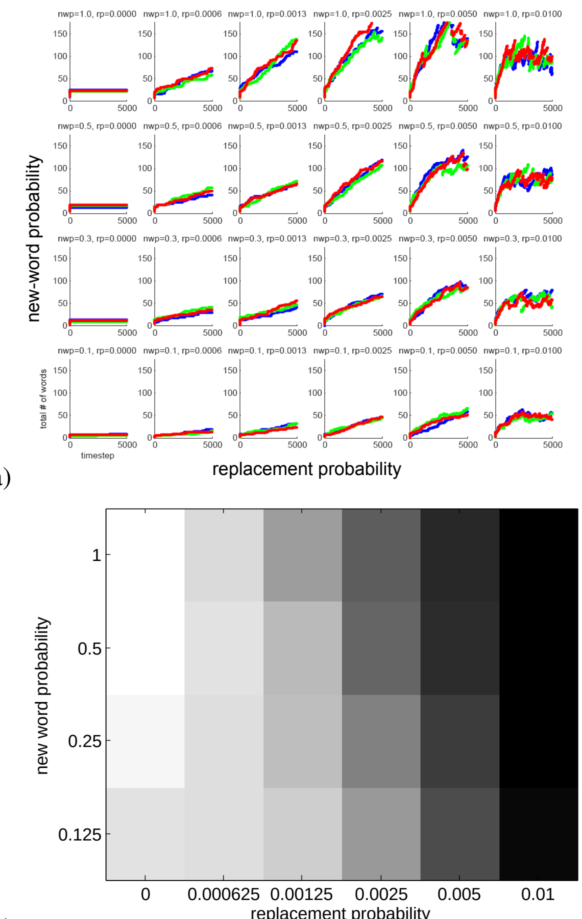


Figure 2: Punctuated equilibria of lexical change. Each color represents the mean value across agents of a word (on one of the dimensions) over time, with a maximum lexicon size of four words. Top: objects are generated uniform-randomly; bottom: objects are generated from four "real" meaning distributions.

# Results

## Stability

As long as new agents are being introduced, the lexicon will continue to change. This effect is caused by the fact that new agents will occasionally see objects for which they have not yet learned a word, so they will create new words that are then adopted into the lexicons of the other agents. The changing lexicon follows a pattern of stable periods punctuated by periods of fluctuation (Figure 2). Qualitatively, the presence of "real" meanings (as opposed to objects with uniform-random feature values) results in less variability during the stable periods but longer periods of wide fluctuation.




(a)

(b)

Figure 3: (a) Total number of words in the population over time, for different probabilities of replacing agents and creating new words. Colors represent three different runs for each set of parameter values. (b) Asymptotic probabilities of understanding (black = chance, white = perfect), for the same parameter values as in (a).

When the agents are allowed an unlimited number of words, the continual creation of new words means that the total number of words in the population will increase indefinitely. Depending on the probability of introducing new words and the probability of replacing agents, the increasing number of words used in the population can be beneficial or not, and the likelihood of successful communication can vary from chance to almost perfect (Figure 3). With an unchanging population (far left column), new words will eventually stop being introduced, and if only a small number of words are created, the agents will maintain a significantly lower probability of understanding each other. On the other hand, when the population is changing very quickly (far right column), there is not enough time for the agents to learn the same lexicon, so their probability of understanding each other is near chance.

The results in the remainder of this section are for simulations with a new-word probability of .1 and an agent-replacement probability of .0005 (or 0 where noted). All results reported are for cross-situational learning unless specified otherwise.

## Division of meaning space

If objects are generated uniform-randomly from the space of features, there is no structure present in the world to shape the formation of language. Nevertheless, the agents will divide the space relatively efficiently given the number of words they are allowed to use. The lexicons they create result in probabilities of understanding that are above chance, and are usually quite close to how often they would understand each other if they divided the space into equal-sized regions with hard boundaries (not probabilistic-distributions). Figure 4 shows the asymptotic level of understanding for each of several maximum lexicon sizes, comparing cross-situational learning versus ostensive learning, and agent-replacing versus no replacement.

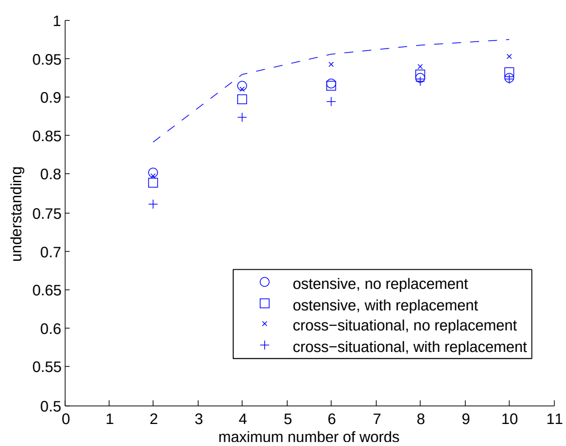



Figure 4: Actual understanding (points) versus optimal understanding (dotted line) for several values of the maximum lexicon size. Results are averages over 10 runs for each parameter setting.

## Synonyms and polysemes

When the agents are allowed a fixed (small) number of words, they will eventually have that many words, which might not be the same as the number of "real" meanings in the world.

If they are allowed more words than the number of meanings they need to cover, they generally do not have words whose meanings are exact duplicates of each other. Instead, they will use some words to cover multiple meanings, or divide a single meaning into subparts with separate words. This is in accordance with the real-life phenomenon of the absence of exact synonyms, as well as the explanation of Nowak et al. (1999) that random sampling of examples will eventually favor one meaning over another.

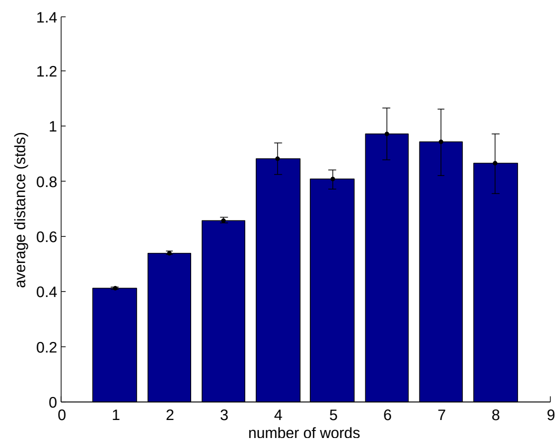



Figure 5: The more words there are to refer to a single meaning (or set of meanings), the farther the words are from the center of the real meaning, indicating that exact synonymy is avoided. Average distance to the center of the real meaning is plotted against number of words for a given meaning, over runs of 1000 timesteps, with a maximum of 10 words. (Error bars represent standard error of the mean, but the samples were not independent.)

If the number of words allowed is fewer than the number of meanings to be covered, there will have to be meanings that share words. When all objects have an equal probability of occurring in the same context, the communicative effectiveness of the lexicon will be the same regardless of which meanings share words. But when certain objects are more likely or less likely to be present together, whether those objects share a word can affect how often the agents understand each other. Simulations using four "real" meanings (objects) with three words found that the more often a pair of objects occurred together, the less often the same word was used to refer to both (Figure 6). In other words, objects were more likely to share a word when the context could disambiguate between them.

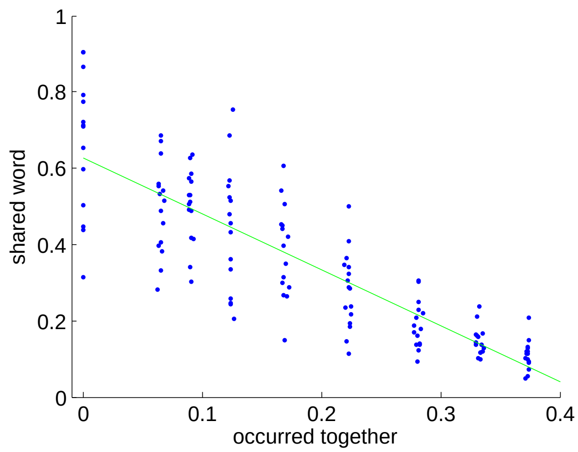



Figure 6: When two objects are more likely to occur in the same context, they are less likely to share the same word. Points shown are 15 runs at each of 9 co-occurrence probabilities for a pair of the four objects. Best-fit line is shown in green.

# Conclusions

This paper demonstrates that in a relatively simple model, it is possible for agents to develop a shared lexicon even when (a) the space of meanings is continuous so there is not a finite set of possible referents, (b) each agent is learning from a different set of examples, and (c) the listener does not always know the speaker's intended referent. Not only do they reach a shared lexicon, but they also efficiently assign words to meanings, both in the presence of a set of "real" meanings and in the case where objects are generated randomly.

The results also suggest how the structure of the world and the behavior of the population affect the dynamics of lexical change. Higher probabilities of replacing agents and adding words cause lower levels of communicative understanding; and the presence of "real" meanings causes less fluctuation during stable periods but more during transitions.

## Future directions

Many additional questions could be explored with this model, such as varying the shapes and positions of the "real" meanings in the space, increasing the dimensionality of the space, changing the number of people, or analyzing the dynamics of the stability of the lexicon over time.

There are also several obvious ways in which the model could be extended.

Meanings are currently represented as uncorrelated Gaussian distributions in each dimension. It would be more realistic to allow other types of distributions, for example distributions with multiple peaks (corresponding to polysemous words whose different meanings do not share any feature values).

There is currently no representation of the phonetic forms of words. If this aspect of communication were incorporated into the model, it could take into account factors such as the length of a word (effort required to produce it) relative to the amount of information it carries and the probability of mishearing it.

The model does not currently include any social structure: each agent is equally likely to talk to each other agent, and the words he uses do not depend on who he is talking to. It would be interesting to make agents whose behavior is different toward other agents depending on their identity. For example, people might be more likely to talk with others of similar ages, and might talk differently to old people than to young people.

## Acknowledgements


I would like to thank Eric Smith for helpful discussion and feedback throughout the course of this project.

This work was conducted through the Research Experiences for Undergraduates program at the Santa Fe Institute, funded by the National Science Foundation.


# References


Gale, W. A., Church, K. W., & Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the workshop on speech and natural language* (p. 233-237). Association for Computational Linguistics.

Khetarpal, N., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

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

Nowak, M. A., Plotkin, J. B., & Krakauer, D. C. (1999). The evolutionary language game. *Journal of Theoretical Biology*, *200*, 147-162.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436-1441.

Roy, D. K. (1999). *Learning words form sights and* *sounds: A computational model*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558-1568.

Steels, L. (1996). Emergent adaptive lexicons. In *From animals to animats 4: Proceedings of the fourth international conference on simulation of adaptive behavior* (p. 562-567). The MIT Press.

Steels, L., & Kaplan, F. (1998). Spontaneous lexicon change. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, *2*, 1243-1250.

Steels, L., & Kaplan, F. (2000). AIBO's first words. the social learning of language and meaning. *Evolution of Communication*, *4*(1), 3-32.

Tenenbaum, J. B., & Xu, F. (2000). Word learning as bayesian inference. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*.

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

Yarowsky, D. (1993). One sense per collocation. In *Proceedings of the workshop on human language technology* (p. 266-271). Association for Computational Linguistics.