5

Probabilistic Association and Denotation in Machine Learning of Natural Language

P. Suppes, L. Liang

We have taken what we believe is a new tack in the approach to machine learning by using in a very explicit way principles of association and generalization derived from classical psychological principles. The principles we used were, however, much more specific and technically developed.

The fundamental role of association as a basis for conditioning is thoroughly recognized in modern neuroscience and is essential to the experimental study of the neuronal activity of a variety of animals. For similar reasons its role is just as central to the learning theory of neural networks, now rapidly developing in many different directions. We have not, however, made explicit use of neural networks, but have worked out our theory of language learning at a higher level of abstraction. In our judgment the difficulties we face need to be solved before a still more detailed theory is developed.

The classical psychological principles of learning used here have been thought by linguists to be wholly inadequate as the basis for a theory of language learning. Nothing could be further from the truth. Skinner's [6] naive formulation of the problems of language learning was rightly attacked by Chomsky [5], but no serious alternative learning theory has been offered by linguists even today.

In the first section we briefly describe our approach to machine learning of natural language. In the second section we focus on the problem of denotation that is important in our use of probabilistic association of words and their meaning. In the third section we outline the backgound cognitive and perceptual assumptions of our machine learning work. In the fourth section we formulate explicitly our two general

COMPUTATIONAL LEARNING AND PROBABILISTIC REASONING
Editor Alex Gammerman

axioms of association and denotation, but do not state the additional axioms describing the full learning process. These may be found in our previous publications, with some changes being made over time [1, 2, 3]. In the fifth section we formulate and analyze two related but distinct denotational learning models. Finally, in the sixth section we present some empirical results.

5.1 Our Approach to Machine Learning

Without going into all the details, we want to convey a rather clear intuitive sense of the process of learning of natural language in terms of the various events that happen when an utterance is given to a robot. (In this and succeeding sections we shall refer to robots, but it should be understood that the basic program of machine learning would apply without serious modification to other applications, for example, machine learning of physics word problems, which we briefly consider later.) Also, following standard learning usage, we shall often speak of trials where, of course, we mean that the trial begins with a command in the form of an utterance to be executed by the robot.

The most important way to describe conceptually the learning process our program embodies is in the description of the state of memory of a robot at the beginning of each trial. There are four aspects of this memory that are changed due to learning. The first is the association relation between words of a given language and internal symbols that have as denotations actions, objects, properties and relations in the robot's world. A central problem is to learn in each language what word is properly associated with a given internal symbol. A second aspect of the memory that changes is the denotational value of a given word, which will affect its probability of being associated. It is this denotational-value aspect of the memory that we describe in some detail, beginning in the next section. The third part that changes is the short-term memory that holds a given verbal command for the period of the trial on which it is effective. This memory content decays and is not available for access after the trial on which a particular command is given. This means that at the beginning of the trial, before a command is given, this short-term buffer is empty. What we have said thus far, could, with some stretching, fit into classical theories of association, but for language learning it is quite evident that the association relation and some simple features of short-term memory are certainly not enough.

The fourth aspect is the important one of learning grammatical forms. Consider the verbal command *Get the nut*. This would be an instance of the grammatical form *A the O*, where *A* is the category of actions and *O* is the category of objects. This form actually represents a mild oversimplification, because we do not have just a single category of actions. There are several subcategories, depending upon the number of arguments required, and certain other natural semantical requirements as well. The example will illustrate how things work, however. The grammatical forms are derived by generalization

only from actual instances of verbal commands given to the robot. No prior knowledge of any sort of the grammar of the natural language to be learned is available to the robot. Also important is the fact that associated with each grammatical form as it arises from generalization are the associations of the words which have been the basis for the generalization, along with their internal representations. For example, if *Get the nut* were the occurrence in which the grammatical form just stated was generated, then also stored with that grammatical form would be the associations *get* $\sim$ *$g* and *nut* $\sim$ *$n*, where *$g* and *$n* are internal symbols whose denotations are known to the robot. When incorrect associations are deleted by further learning, the grammatical forms based on such associations are also deleted.

5.2 Problem of Denotation

In the probabilistic theory of machine learning of natural language which we have been developing, we have encountered in a new form a standard problem in the analysis of the semantics of natural language, namely, how to handle words that are nondenoting. We do not mean nondenoting in some absolute sense, but relative to a fixed set of semantic categories. These categories in the robotic case are, roughly speaking, the categories of actions, objects, properties and relations. It may well be that in some elaborate set-theoretical semantics of natural language, nondenoting words like the definite article *the* denote a complicated set-theoretical function, but the relevance of such an elaborate semantics to language learning is doubtful. In the robotic context, we have something simpler and closer to the common man's view of what denotations are. We take as denoting words color and object words, common nouns, familiar concrete action words, *etc.*. We take ordinary prepositions in English and sometimes other devices in other languages to denote relations in most cases. On the other hand, our computational semantics centered on physics words problems, with an internal equational language of physical quantities, is further removed from common sense ideas.

When a child learning a first language or an older person learning a second language first encounters utterances in that new language, there is no uniform way in which nondenoting words are marked. There is some evidence that various prosodic features are used in English and other languages to help the child. For example, in many utterances addressed to very young children, the definite or indefinite article is not stressed but rather the common noun it modifies, as in the expression *Hand me the cup*. But such devices do not seem uniform and in any case are not naturally available to us in our machine-learning research, where we use written input of words without additional prosodic notation.

As has already been made clear, a central feature of our approach to machine learning is the probabilistic association between words of the natural

language being learned and denoting symbols of the internal language. It is appropriate that at the beginning all words are treated equally, and so the associations are formed from sampling based on a uniform distribution. On the other hand, after many words have been learned and a good deal of language has been acquired by the robot, it is very unnatural, and also inefficient, if the robot is now given, for example, the esoteric command *Get the astrolabe*, to have the internal symbol *$ast* be associated with equal probability with the definite article *the* and *astrolabe* — we assume here that the association of *get* is already correctly fixed. After much experience, what we want is that there is very little chance of associating the definite article *the* with any denoting symbol.

To incorporate such learning many variant models are easily formulated. We have restricted ourselves to two which bring out the most salient distinction to consider. Should the denotational value of a word change only when it occurs in an utterance that is responded to incorrectly or not at all, or should the denotational value change every time it occurs regardless of whether a correct response is given? Model I is of the first type — learning only from errors. Model II is of the second type — learning occurs on all trials.

These two types of models have a long history of application in the study of human and animal learning. In such studies the empirical question centres on which kind of model approximates more closely the learning that is taking place in the organism. In machine learning this question is answered by fiat, that is, by implementing in the machine-learning program one of the models. But empirical questions remain. Having implemented a model, we study how well it works in providing dynamically changing denotational values for words of a given language. In particular, we analyze in detail, by extensive computation, the error rates of association arising for different parameter values in the different models and for different languages. In this chapter, we present data for English and Chinese.

5.3 Background Cognitive and Perceptual Assumptions

Before explicitly formulating the learning principles of association and denotation we use, we first state informally assumptions we make about the cognitive and perceptual capacities of the class of robots, albeit as yet quite limited, we work with.

1. *Internal language.* The robot has a fully developed internal language, which it does not learn. It is technically important, but not conceptually fundamental, that in our case this language is LISP. When we speak here of the internal language we refer only to the language of the internal representation, which is itself a language at a higher level of abstraction, relative to the

concrete movements and perceptions of the robot. It is the language of the internal representations held in memory that provides the direct interface to the natural-language learning. In fact, most of the machine learning of a given natural language can take place through simulation of the robot's behaviour by using just the language of the internal representation. The first associations learned are between the internal representation in memory of a coerced action and a contiguous verbal utterance in the natural language being learned.

The fundamental importance of this internal representation of a coerced action can be recognized by considering a parallel case of animal learning. When a dog is trained to *Get the paper* or *Get the ball* by being led through the desired action or by some related technique, the residue in memory of what we term the coerced action is surely drastically abstracted from the perceptually rich context of the demonstrated action desired, and it is that abstracted internal representation in memory that must be associated to the verbal stimulus in order for the dog later to perform the desired action upon hearing the verbal command. We are a long way from knowing even the general structure of the dog's internal representation in memory of the action. In this limited sense, life with a robot is much easier, for we ourselves create the form of its internal representation.

2. *Objects, relations and properties.* We furthermore assume the robot begins its natural-language learning with all the basic cognitive and perceptual concepts it will have. In other words, our first-language learning experiments are pure language learning. Any learning of new concepts is delayed to another phase. For example, we have assumed that the spatial relations frequently referred to in all, or at least all the languages we consider in detail, are already known to the robot. This is quite contrary to human language learning. For example, probably in no widely used natural language at least, do children at the age of thirty-six months use or fully understand the relations of left and right. To avoid misunderstanding, we emphasize that we consider it an important future task to have the robot also learn the familiar spatial and temporal relations.

3. *Actions.* What was just said about objects and relations applies also to actions, represented in English by such verbs as *pick up, get, place, screw, etc.*. The English, of course, must be learned, but not the underlying actions.

4. *Associations and grammatical forms.* Before stating any formal principles of learning, we feel it is desirable to describe as informally and intuitively as possible the learning setup we use. Consider the English command *Pick up the screw*, no part of which has as yet been learned by the robot. The learning steps may be roughly schematized as follows:

(i) By coercion, or simulation of coercion, the robot creates in memory an internal representation of the coerced action of picking up the screw;

For statement of learning principles we show this internal represen-

tation, not as a LISP expression, but just as a schematic function *I(...)* of the denoting terms in the LISP expression. Here, by *denoting terms* we mean the names in the internal language of the actions, objects, properties and relations mentioned. The internal representation of *Pick up the screw* is then *I($p,$u,$s)*, where *$p* = the action of picking, *$u* = the direction up and *$s* = screw.

(ii) By contiguity the robot associates the verbal utterance and the internal representation

$$Pick\ up\ the\ screw \sim I(\$p,\$u,\$s),$$

where $\sim$ is the symbol we use for association;

(iii) By probabilistic association, the robot associates the internal denotations with the English words, with one possibility the following incorrect result:

$$pick \sim \$s,\ up \sim \$p,\ screw \sim \$u.$$

We need to observe the following:

a. We assume from the beginning the robot knows word boundaries, as delineated by the typed input. This is an example of an assumption that is natural for robots, but clearly false for very young children;

b. For our simple example, there are 24 possible ways of associating the three internal symbols to the four denoting words in the English utterance. We initially assign to each of these 24 possibilities equal probability, but as trials contine, modify the probability by dynamic changes in denotational values, as is explained later in detail.

(iv) After the associations are made, by the principle of generalization, which we call the category generalization, each word is assigned the category of its associated internal symbol. In the present case *pick* $\in$ *O* — the category of objects, *up* $\in$ *A* — the category of actions and *screw* $\in$ *R* — the category of relations. A grammatical form is then also generalized from the verbal command:

$$O\ \ A \text{ the } R$$

which, like the assigned categories is wrong for English, but remember that this is just the starting point of learning. With this grammatical form is associated its internal representation *I(A, R, O)* which characterizes its meaning.

(v) A new command is presented as the next step, say *Pick up the nut.* By coercion the internal representation *I($p, $u, $n)* is created (see (i) above). The robot then first searches its memory to see if any of

the words uttered are associated to one of the internal denotations. Here the result is *up* $\sim$ *$p*, and also the classification of *the* as a nondenoting word is found. There are then six possibilities of probabilistic association for *pick*, *the* and *nut*. Note that the earlier incorrect association of *pick* with *$s* does not appear here, which means that at this stage of learning it will be changed. So, let us suppose the new associations are

$$pick \sim \$u,\ nut \sim \$n.$$

We also have as a new grammatical form

$$R\ A\ the\ O$$

which though incorrect, now has only the confusion of the associations of *pick* and *up* as its source. To correct these associations we must separate the constant pairing of *pick* and *up*, which is what we do. In any case, we form at once the association to the internal representation:

$$R\ A\ \text{the}\ O \sim I(A,\ R,\ O).$$

(vi) Learning stops whenever the following steps of interpretation can be successfully completed upon giving the robot a verbal command:

a. An association to an internal denotation or a nondenoting classification is found in memory for each word;

b. The category of each word is found in memory;

c. The grammatical form resulting from (b) is found with an associated internal representation in memory;

d. The command is correctly executed on the basis of the internal representation.

5.4 The General Axioms of Association and Denotation

We state the axioms in a general form, to be made more specific later, but we assume already that each word a of the target natural language has a denotational value $d_n(a)$ on each trial. This value changes from trial to trial according to the two different models presented in the next section.

I. *Probabilistic association.* On any trial n, let a natural language sentence s be associated to σ, its internal representation, let $\{a_i\}$ be the set of words of s not associated to any internal denoting symbol of σ, let $d_n(a_i)$ be the current denotational value of each such a_i and let $\{\alpha_j\}$ be the set of internal denoting symbols not currently associated with any word of s. Then:

(i) an element α_j is uniformly sampled without replacement from $\{\alpha_j\}$;

(ii) at the same time an element a_i is sampled without replacement from $\{a_i\}$ with the sampling probability

$$p_n(a_i) = \frac{d_n(a_i)}{\sum_{\{a_i\}} d_n(a_i)};$$

(iii) the sampled pairs are associated, *i.e.* $a_i \sim \alpha_j$;

(iv) sampling continues until either the set $\{a_i\}$ or the set $\{\alpha_j\}$ is empty.

II. *Denotational value computation.* If at the end of a trial a word a in the presented sentence is associated with some internal symbol α, then $d(a)$, the denotational value of a, increases and if a is not so associated $d(a)$ decreases. Moreover, if a word a does not occur on a trial, then $d(a)$ stays the same unless the association of a to an internal symbol α is broken on the trial, in which case $d(a)$ decreases.

5.5 Denotational Learning Models

We now turn to the description of the two models we consider. We begin with Model I — learning denotational values only from errors. Only two axioms are needed, but the exact description of the conditions for application of the learning operators are rather complicated. We describe these conditions informally rather than introduce a great deal of formal notation that we subsequently would not make much use of.

Model I

1. *If on trial n a wrong response or no response is given to the verbal command uttered at the beginning of the trial, and if after the coercion and probabilistic association of some (perhaps all) words of s to internal symbols of the internal representation σ of the coerced action, a word a_i is now associated to an internal symbol α_j of σ, i.e. $a_i \sim \alpha_j$, then*

$$d_{n+1}(a_i) = (1-\theta)d_n(a_i) + \theta,$$

and if a_i is not so associated with any denoting internal symbol

$$d_{n+1}(a_i) = (1-\theta)d_n(a_i),$$

where the learning parameter θ satisfies the constraint $0 < \theta \leq 1$, and so does the initial value $d_1(a_i)$, which is the same for all words a_i.

2. If on trial n a correct response is given to the verbal command s, then no denotational values of words are changed, i.e. for all words a_i

$$d_{n+1}(a_i) = d_n(a_i). \tag{5.1}$$

Moreover, if a word a_i does not occur on trial n, equation (5.1) also holds unless the association of a_i to an internal symbol α_j is broken on trial n, in which case

$$d_{n+1}(a_i) = (1 - \theta)d_n(a_i).$$

There is much to be said for learning only when errors are made. But learning is not a monolithic phenomenon. Many kinds of skills improve with practice even after explicit errors are no longer made. Children, for example, continue to learn to read faster or to solve elementary mathematical problems faster long after their error rates are insignificant. In these and similar cases the evidence is substantial that learning, as reflected in speed of response, is taking place on trials with correct responses as well as on those with errors. The axioms for Model II reflect this kind of learning.

Model II

If, at the end of trial n, a word a_i in the presented verbal stimulus is associated with some denoting internal symbol α_j of the internal representation σ of s at the end of the trial, then

$$d_{n+1}(a_i) = (1 - \theta)d_n(a_i) + \theta,$$

and if a_i is not so associated,

$$d_{n+1}(a_i) = (1 - \theta)d_n(a_i).$$

Moreover, if a word a_i does not occur on trial n, then

$$d_{n+1}(a_i) = d_n(a_i),$$

unless the association of a_i to an internal symbol α_j is broken on trial n, in which case

$$d_{n+1}(a_i) = (1 - \theta)d_n(a_i).$$

It should be apparent that Models I and II have the same two parameters, namely, the initial denotational value d_1 for all words of a language and the learning parameter θ. The general empirical question concerns what values of the parameters work best for each model for a given language, or more explicitly, for a given finite sample of utterances of the language, the one used for learning. Of course, there is no unique best criterion of parameter performance. We have used the following, although many others are easy to formulate.

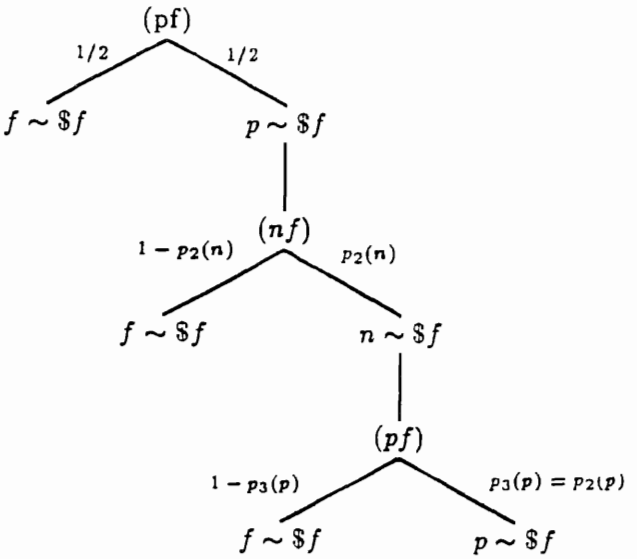

Figure 5.1: Initial part of the tree for language L_3.

First, to introduce some notation, useful now and later, let

$E_\delta(d_n)$ = the expectation or mean of d_n on trial n for all denoting words (indicated by the subscript δ),

$E_\nu(d_n)$ = the expectation of d_n on trial n for all nondenoting words.

The criterion we use is the first trial m^* on which for the finite sample used

$$E_\nu(d_{m^*})/E_\delta(d_{m^*}) < \epsilon. \tag{5.2}$$

In most of the analysis reported below we have chosen $\epsilon = 0.01$.

It is easy to see that if we classify externally words as denoting or nondenoting, as we naturally do on learning trials, then in simple examples m^* may not exist. Consider the language L_3 with three words and two commands, *Please forward!* and *Now forward!* with *please* and *now* being nondenoting and *forward* having the same denotation as *$f*. Let the learning parameter $\theta = 1$. Then suppose on trial 1 with the utterance *please forward* given at the beginning of the trial, as the result of sampling *please* is associated with *$f*, and *forward* has no association. Then with $\theta = 1$ we have at the end of this first trial, ready for trial 2, $d_2(please) = 1$, $d_2(forward) = 0$, $d_2(now) = d_1(now)$. Then on the presentation of (*now forward*) at the beginning of the next trial, the internal symbol *$f* is associated with neither *now* nor *forward*. After coercion, the probability of association with *now* is

$$\frac{d_2(now)}{d_2(now) + d_2(forward)} = \frac{1}{1+0} = 1,$$

and obviously $prob_2$ (*forward* ~ *$f*) = *0*. At the end of this trial, trial 2, $d_3(please) = d_2(please) = 1$, $d_3(forward) = d_2(forward) = 0$, $d_3(now) = 1$.

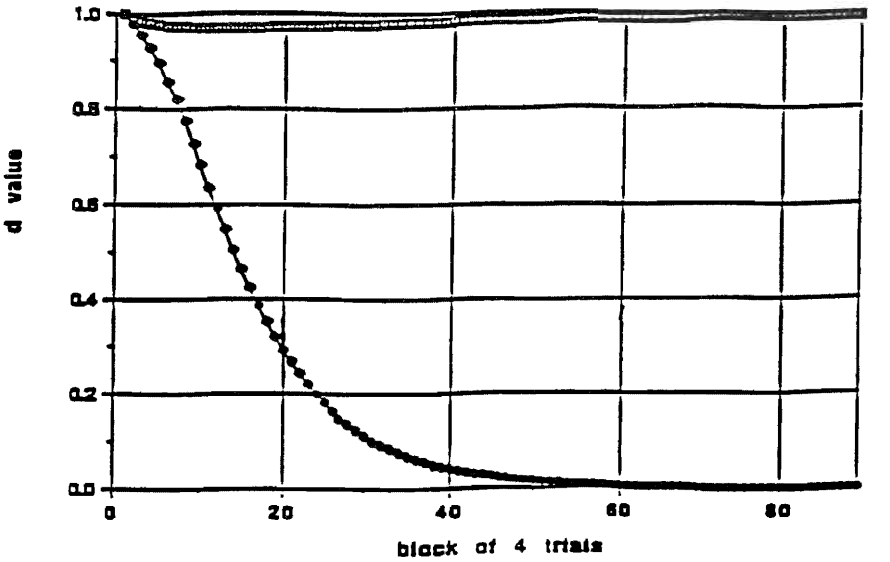

Figure 5.2: Mean denotational learning curves for English robotic corpus.

From here on these values of d_n are constant, *i.e.* for $n \geq 3$, $d_n(please) = d_n(now) = 1, d_n(forward) = 0$. (This computation holds in both Models I and II.) What is also interesting about this example is that if we define denoting and nondenoting internally just in terms of the values of d_n for various words, with a word a_i denoting whenever $d_n(a_i) \geq 1 - \epsilon$ and nondenoting whenever $d_n(a_i) \leq \epsilon$, then in this example $m^* = 3$ for any $\epsilon > 0$.

Asymptotic behaviour. We can study analytically the asymptotic behaviour of various denotational learning models only for trivial languages. Even small fragments of several hundred utterances of a natural language can only be studied numerically from the standpoint of denotational learning. Nevertheless, analysis of trivial examples gives some insight into how the models behave.

We first examine the asymptotic behaviour of Learning Model I for language L_3. Figure 5.1 shows the initial part of the tree. We can focus our analysis entirely on the right-most branch of the tree, for all other branches have after a finite number of trials the correct association.

For simplicity of computation, let

$$d_1(now) = d_1(please) = d_1(forward) = 1. \tag{5.3}$$

Then it is easy to see that the probability p_r of the right-most branch is the infinite product:

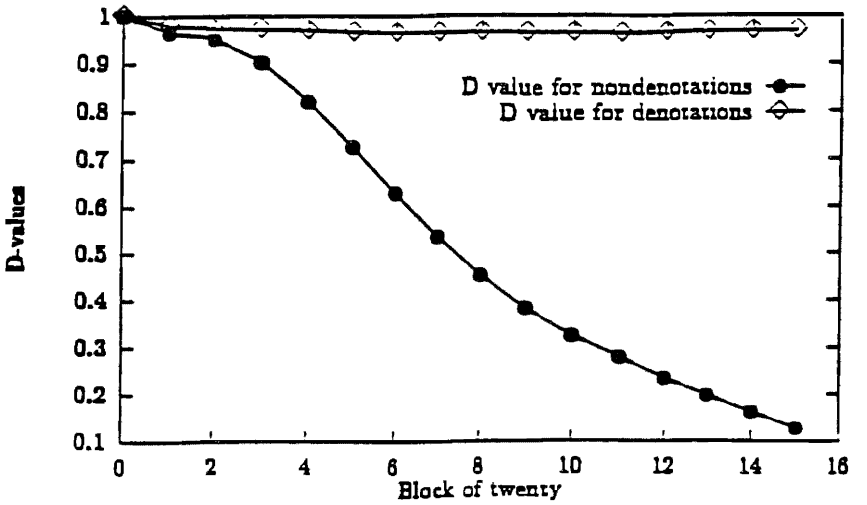

Figure 5.3: Mean denotational learning curves for 60 physics word problems in English.

$$p_r = \frac{1}{2} \cdot \frac{1}{1+(1-\theta)} \cdot \frac{(1-\theta)+\theta}{(1-\theta)+\theta+(1-\theta)^2} \cdot \frac{(1-\theta)+\theta}{(1-\theta)+\theta+(1-\theta)^3} \cdots$$
$$= \frac{1}{2} \cdot \frac{1}{1+(1-\theta)} \cdot \frac{1}{1+(1-\theta)^2} \cdot \frac{1}{1+(1-\theta)^3} \cdots \quad . \tag{5.4}$$

Let $\alpha = (1-\theta)$. It is most direct to study the behaviour of $1/p_r$.

$$\frac{1}{p_r} = 2 \cdot (1+\alpha) \cdot (1+\alpha^2) \cdot (1+\alpha^3) \cdots \tag{5.5}$$

For $0 < \alpha < 1$, $\sum \alpha^n$ converges to $\frac{1}{1-\alpha} = \frac{1}{\theta}$.
But this convergence is a necessary and sufficient condition for the convergence of (5.4) to a positive real number, say, $c > 0$. So $p_r = \frac{1}{c} > 0$.

Without giving a detailed analysis, we remark that quite similar results are obtained from a different, but familiar class of learning models, ones with commuting operators. Here is a generic example:

$$d_{n+1}(w) = \begin{cases} \alpha\, d_n(w) & \text{if word } w \text{ occurs on trial } n \text{ and is not associated with any internal symbol,} \\ \alpha\, d_n(w) & \text{if } w \text{ does not occur on trial } n \text{ but its association is broken,} \\ \beta\, d_n(w) & \text{if } w \text{ occurs on trial } n \text{ and is associated with an internal symbol,} \end{cases}$$

where $\alpha < 1$ and $\beta \geq 1$. We do not pursue these models here.

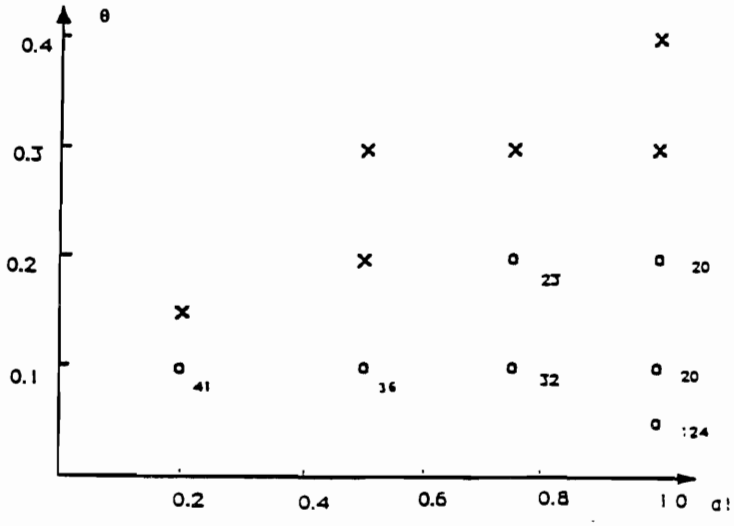

Figure 5.4: Rate of denotational learning with different values of $d_1(a_i)$ and θ.

5.6 Some Empirical Results

In Figure 5.2, we show the mean denotational learning curves of Model II for the English robotic corpus of approximately 400 commands. The only two nondenoting words in the corpus were the definite and indefinite articles *the* and *a*. The bottom curve is for these nondenoting words, and the upper curve for the denoting words. The parameters were set at $d_1(a_i) = 1$ and $\theta = 0.03$. Using the same parameters the results were similar, but with slower learning, for the corresponding Chinese robotic corpus which had eight nondenoting words. (For more linguistic details, see [4].

In Figure 5.3, we show the corresponding mean curves for 60 physics word problems in English, with again the only two nondenoting words being *the* and *a*. The internal language is completely different from the robotic case, but we will not attempt to describe it here, except to say that it is essentially a pure equational language for physical quantities. We emphasize, however, that the learning axioms were the same for the robotic commands and the physics word problems.

Finally, in Figure 5.4 we show the effects of choosing different initial values $d_1(a_i)$ and simultaneously different values of θ in Model II for the robotic English corpus. The criterion of learning was that the ratio (5.2) be such that $\epsilon < 0.01$. The number of trials to achieve this result is printed next to each data point.

5.7 Conclusion

In spite of the problems still to be solved in our theory of machine learning of natural language, prospects for use are not only a distant hope. We list three.

First, in the relatively near future real applications to well-defined domains of activity and their relevant sublanguages will appear. Some of the earliest successful technical examples are likely to be in medicine, from the emergency room to the office dictation of medical records.

Second, it is likely, even if not anything like certain, that within the next decade oral communication with computers, as with people, will be the most important and most used form of communication. If so, single words and phrases will not be good enough. A rich natural sublanguage will be used, and computers will need to learn it. No doubt, the progress on comprehension may be faster than on production in the early years.

Third, important applications early in the next century at least will be in two domains, reflected in the early work reported here. There will be robots that talk and, above all, listen to instructions, and immobile computer-tutors that also talk and listen, and in the process teach the way a good tutor should.

References

[1] Suppes P., Liang L. and Böettner M. (1992) Complexity Issues in Robotic Machine Learning of Natural Language. In Lam L. and Naroditsky V. (eds) *Modelling Complex Phenomena*, Springer-Verlag, New York, 102–127.

[2] Suppes P., Böettner M. and Liang L. (1995) Comprehension Grammars Generated from Machine Learning of Natural Languages. *Machine Learning* 19: 133–152.

[3] Suppes P., Böettner M., Liang L. and Raymond R. (1995) Machine Learning of Natural Language: Problems and Prospects. *Proceedings of the Second World Conference on the Fundamentals of Artificial Intelligence*, France, 3–7 July 1995, 511–525.

[4] Suppes P., Böettner M. and Liang, L. Machine Learning Comprehension Grammars for Ten Languages. To appear.

[5] Chomsky, N. (1959) Review of B. F. Skinner *Verbal Behavior*. *Language* 35, 26-58.

[6] Skinner, B. F. (1959) *Verbal Behavior*. New York: Appleton.