21

LEARNING FROM POSITIVE-ONLY EXAMPLES:

The Subset Principle and Three Case Studies

Robert C. Berwick
Massachusetts Institute of Technology

Abstract

A key issue for learning theory has been the relative importance of domain-independent learning versus domain-specific learning. How much of learning is attributable to general learning methods like inductive generalization, and how much is attributable to particular techniques and representations that apply only to specific domains? This chapter explores this issue through an analysis of the role of one general principle in the context of several very specific domains. Angluin (1978) established a necessary and sufficient condition for the acquisition of a (recursive) language (in the sense used by Gold, 1967) from positive-only evidence. The effect of this condition is to impose an ordering on possible sequences of guesses about the target language combined with the array of possible data sequences in such a way that the acquisition system always guesses the narrowest possible language compatible with the data given so far. In this chapter, this "Subset Principle" is applied to three areas in which extensive domain-specific knowledge is present. First, concept acquisition in children, as studied by Keil (1979), is considered, followed by an examination of the area of natural language, where syntactic constructions and phonological distinctive feature systems are explored. Constraints on ιhese systems are shown to follow the Subset Principle, providing some evidence that natural learning systems are designed to be easily learnable.

21.1 INTRODUCTION

What are the scope and power of "general" learning principles? General methods like inductive generalization or analogy have figured prominently in any discussion of learning. Those who have embraced such general principles range across the scientific spectrum, from Skinner and Piaget to Newell and Simon. More recently, however, the very existence of general learning principles has been questioned:

> What I intend to signify in referring to the doctrine of uniformity of mind is . . . that there are general principles of learning that underlie all of these systems, accounting for their development: "multipurpose learning strategies," as they are called, that apply "across the board." In contrast, it might be proposed that various "mental organs" develop in specific ways . . . and that multipurpose learning strategies are no more likely to exist than general principles of "growth of organs" that account for the shape, structure, and function of the kidney, the liver, the heart, the visual system, and so forth. (Chomsky, 1980, 245)

Chomsky claims that there are no general principles of human learning. Each domain, be it language, motor control, vision, or mathematics, has its own particular constraints. Even if this were true, however, there would still be room for a domain-independent learning theory; it would consist of general principles supplementing the particular constraints of each domain.

As a simple example, consider Winston's classic program that learned the descriptions of scenes made from toy blocks, such as *arch* or *house* (Winston, 1975). Here the domain-dependent constraints include those of the *representation language* primitives, that is, the basic vocabulary used to describe blocks world scenes, such as the predicates *touch* or *is-a-brick*, and the way that these basic predicates can be pasted together. The presumably domain-independent principles of the blocks world include generalization heuristics such as "require link," a descriptor-modification introduced when a negative example shown by a teacher establishes that a particular feature of a block model *must* be present. For instance, if Winston's program was shown a nonarch with its top (or lintel) lying on the ground, it could apply the require link heuristic to the descriptor that the two arch columns *must* support the lintel. The require link heuristic is part of general learning theory.

The question still remains of the relative contribution of general learning theory to the explanation of human learning; that is, can it help us account for observed patterns of human cognitive development? The answer is often assumed to be yes, but matters are far from clear. The aim of this chapter is to show that there is a viable general learning theory that helps explain why human learning proceeds the way it does. It will be shown that there is a powerful constraint on the order in which a learner should consider hypotheses, called here the *Subset Principle.* Armed with this principle, the author then shows how to explain some observed patterns in human learning in the domains of language and concept acquisition.

To see just why the explanatory power of general learning theory can be questioned at all, let us consider the example of language. Modern linguists generally make the idealization that language acquisition is "instantaneous"—that is, they imagine that the language learner is presented with all the "input data" (sentences of their language to learn) at once. This is clearly false. Children hear sentences strung out over time. Linguists know that this idealization is false, just as physicists know that there are no frictionless planes. Yet the idealization has proved remarkably successful: so far, no generalizations about the properties of natural languages have been lost because of the idealization of instantaneous acquisition. At least the linguists would claim that this is true. In fact, it is widely assumed that there is no theory of language learning at all. John Marshall has put things this way:

> There is, however, a very general problem with practically all studies of language development, whether investigated from the standpoint of rule acquisition, strategy change, or elaboration of mechanism. The problem arises both for accounts that postulate "stages" of development (i.e., a finite number of qualitatively distinct levels of organization through which the organism passes en route from molecule to maturity) and for accounts that view development as a continuous function of simple accumulation. The difficulty is this: No one has seriously attempted to specify a mechanism that "drives" language acquisition through its "stages" or along its continuous function. Or more succinctly: there is no known learning theory for language. (Marshall, 1979, 443)

This chapter provides just such a theory. The Subset Principle heuristic "drives" language acquisition through its stages. It is not a constraint particular to language, just as the require link heuristic is not particular to the domain of toy blocks. What is the Subset Principle? Intuitively, it is a strategy of "timid acquisition": If possible guesses can be arranged in a subset relationship, then the learner should make the smallest possible guess about what it should learn consistent with the evidence it has seen so far. This is an exceedingly simple idea, yet it is quite powerful. As it happens, this constraint is necessary and sufficient for successful acquisition given only positive training examples, where successful acquisition is defined as convergence to the correct target description or language after some finite number of training examples. (Recall that a positive training example is an example of the concept to be learned. In the Winston toy blocks world, if one is learning about arches, then a positive example is an example of an arch. A negative example is an example of a nonarch.)

This chapter will focus on two natural learning systems, language and concept development. The Subset Principle can account for a wide variety of constraints in these systems. In the case of concept development, the way that children learn about what kinds of things there are in the world will be examined. The analysis is based on work by Keil (1979). It will be seen that the Subset Principle actually explains the developmental stages that children go through as they learn. Language will be considered next, and two subareas will be examined. The first is phonology, that part of linguistics that studies the inventory of sounds in a language. Linguists know that the possible sounds in a given natural language are actually quite limited—out of fifty or so possible sounds, there will be at most a handful of vowels (like *a* or *u*) and a few dozen consonants (like *p, t,* or *k*). Many languages have far fewer consonants. The vast majority of possible combinations of consonants and vowels is never found. For example, no natural language lacks so-called voiceless consonants (consonants pronounced without the vocal cords vibrating, as in a hissed *s*). Why is this so? The Subset Principle explains why: the gaps are an artifact of timid acquisition. The second subarea of language examined here is syntax. Here too the Subset Principle accounts for a wide variety of otherwise inexplicable constraints.

Finally, the Subset Principle has had two "practical" applications. First, it has been used in a computer model for language acquisition (Berwick, 1980, 1982). This model has successfully acquired a large complement of rules for analyzing English sentences and is now being extended to Chinese and German. Second, recent psycholinguistic evidence has probed for evidence of the Subset Principle in young children's acquisition of syntax; preliminary results confirm the principle.

Before applications of the Subset Principle in specific learning situations are considered, the principle will be described in the abstract.

21.2 THE SUBSET PRINCIPLE

The Subset Principle is actually quite simple. The intuition behind it will be presented first. Let us use the Winston blocks world setting as an example of a typical learning situation. We assume that the learner has at its disposal a fixed representation language with which to describe observed scenes of blocks. The learner is presented with examples and nonexamples of some target concept to learn, such as *arch*. In the case of language, the target concept is the rule system of the language itself, such as English or German. Acquisition proceeds via the presentation of a sequence of positive and negative examples. After each example the learner may make some response and change its current model of the target concept or language. A change is prompted by some difference between the current model and an example. If after some finite sequence of presentations the learner does not change its model and has settled on the correct target model or language, we say that acquisition has succeeded.

Consider how this works in the toy blocks world. If the system's current model of an arch includes two columns supporting a wedge, and if it now receives as an example of an arch a set of blocks with two columns supporting a rectangular brick, then the difference prompts a generalization. Perhaps the top of the arch can be any prismatic solid. Positive examples (examples of arches) induce generalizations. They rule out descriptions that are too specific. On the other side, negative examples rule out certain overgeneralizations. If we present as an example of a nonarch two blocks that touch each other (so that there is no hole between them) plus a wedge on top, then the discrepancy between this and the current model forces the learner to add a *must-not-touch* or *must have a hole between* descriptor to the properties of the two supporting columns. In Mitchell's version space framework (1978), positive examples force the boundary of the "specific descriptions" frontier toward more general descriptions, and negative examples force the boundary of the "maximally general descriptions" toward more specific descriptions.

An important variant of this learning situation restricts the learner to positive examples. This is a crucial assumption for models of language acquisition, where the existence of negative evidence is problematic. Children do not seem to learn their native language through explicit, Berlitz-like training sequences (see Brown and Hanlon, 1970; Wexler and Culicover, 1980). No one tells them that "sentence X is *not* a sentence in English," corresponding to the "this is *not* an arch" examples in the toy blocks world.

Unfortunately, a restriction to positive evidence makes learning harder. The danger is overgeneralization. Suppose a learning program gets only positive examples of a concept or a language. If the program's model of the concept becomes too general, no further positive evidence can dislodge it from its incorrect perch. This is simply because there can be no inconsistency between a too-general model and a positive example. We have no negative examples that tell us that we have gone beyond the correct target description. Remember, it is the negative examples in Mitchell's framework that tell us when we have overgeneralized. In the arch example, if we are limited to examples of arches, and if after seeing two columns supporting a brick we generalize to say that the columns may or may not touch, then no further examples where the columns do not touch will disagree with our description.

A solution is to avoid ever hypothesizing an overly general description. We should *order* our hypotheses in such a way that at each step we are guaranteed never to have formed too general a description. This way, if our description is incorrect, say, too specific, then a later positive example will correct it. If we say that arches can have only wedges on top, then an arch example with a brick on top tells us that we are wrong. If our description is just right and arches always have wedges on top, then we simply never change our original hypothesis. In a word, we want our hypotheses to be *maximally disconfirmable*. In the arch example, again, we see that the right choice to make after seeing an arch where the two columns do not touch each other is to use the descriptor *must-not-touch* to begin with. This description is refutable by a positive example if we are wrong: if someone shows us an arch with the two supporting columns touching, then we change to *may-or-may-not-touch*. In this case, of course, our original description is correct. Note how a description of *may-or-may-not-touch*" is wrong to begin with, since it cannot be disconfirmed by positive examples if the correct target descriptions is *may not touch.*

|| Popper

Consider what it means to have a hypothesis that can be disconfirmed by positive examples. Call M_{hypo} the hypothesized model (like the arch description). Let $\mathbf{M}_{hypo}$ be the set of arches describable by M_{hypo}. Similarly, let M_{true} and $\mathbf{M}_{true}$ be the true description and set of arches so describable, respectively. If M_{hypo} is wrong, then for it to be disconfirmable by positive examples, some example in $\mathbf{M}_{\text{true}}$ should show this; that is, there should be some finite set of examples that is not covered by M_{hypo}. This is just like the brick-topped arch example that tells us that our description of arches as all wedge-topped is wrong. Mathematically, then, we can disconfirm our hypothesis if $\mathbf{M}_{\text{true}}$ is *not* a proper subset of $\mathbf{M}_{\text{hypo}}$. Disconfirmation fails only if any example covered by M_{true} is also covered by M_{hypo}. But this is just to say that disconfirmation fails if $\mathbf{M}_{\text{true}}$ *is* a subset of $\mathbf{M}_{\text{hypo}}$. In this case, M_{hypo} is too general, unless it happens to be exactly correct.

Of course, in any actual learning situation we do not know what the true description is. However, we can still arrange for a sequence of hypotheses to meet the condition of disconfirmability. Consider any two hypotheses h_i and h_{i+1}, where the subscripts mean that the hypothesis $i + 1$ is put forth after hypothesis i. Once again, hypothesis $i + 1$ should *not* be a proper subset of hypothesis i. If we guess first that column supports for an arch *may-or-may-not-touch* then we cannot disconfirm this, since the hypothesis *may-touch* is a proper subset of this first guess. Reversing the logic, what we should do is order our guesses so that each pair $(i, i + 1)$ is disconfirmable. We say that this arrangement of hypotheses satisfies the Subset Principle.

It is not hard to see why the Subset Principle is dubbed "timid acquisition." If a learning system follows the principle, then it most often makes the smallest generalization possible at any given step. (It need not be so timid if it can be certain of receiving disconfirming evidence at some later point. For example, suppose the learner had to choose between two languages, $L_1 = \{a^i, i \text{ is odd}\} + \{a, a^2, \ldots, a^{10}\}$ and $L_2 = \{a^i, i \text{ is even}\} + \{a, a^2, \ldots, a^{10}\}$. Then just a single positive example, say, a^3, can spark an inductive leap to the guess of L_1, even though this may be wrong. For if the learner is wrong, it will eventually get an example such as a^{12} that will prove this to be so. On the other hand, a timid generalizer would not be able to make such a leap.)

It is also not hard to see that the Subset Principle is sufficient to guarantee successful acquisition after some finite number of positive examples. (We might not know how many examples this would take, however.) This is because any step at all is a step in the right direction, as stimulated by a positive example. After some finite number of steps, the system must guess the correct target language or description.

More strikingly, Angluin (1978) has shown that this principle is actually necessary for acquisition given positive-only evidence. Angluin proves this result using the techniques of recursive function theory. In her framework, the "hypotheses" are a family of languages, $\mathcal{L} = \{L_1, L_2, \ldots, L_j, \ldots, L_i, \ldots L_n\}$. The "examples" are finite collections of positive examples, perhaps singleton sets or perhaps not, defined as T_i, where T_i is a positive set of examples for hypothesis L_i. Finally, Angluin uses the term *identifiable* to mean that after some finite number of positive example presentations, the learning procedure (1) guesses the right target language and (2) never changes its guess after this. This is Gold's (1967) traditional definition of identifiability in the limit. The theorem proved is the following:

Theorem 1: Given a family of languages $\mathcal{L}$, $\mathcal{L}$ is identifiable from positive-only evidence if and only if for each target language L_i in $\mathcal{L}$ there exists a computable procedure that enumerates finite sets T_1, $T_2, \ldots$, such that

(1) $T_i \subseteq L_i \subseteq \mathcal{L}$ and

(2) For all $j > i$, if $T_i \subseteq L_j$, then L_j is not a proper subset of L_i.

The Subset Principle is a very general and abstract restriction on acquisition using positive examples. Suppose that hypotheses may be nested, so that each one completely covers the next. This is usually the case in natural languages, when one has a rich theoretical vocabulary and is trying to find a correct description of some target grammar. In this situation, Angluin's result says that one must find the smallest generalization covering the samples seen so far or otherwise one risks overgeneralizing. This kind of problem is also discussed in a more general context by Diettrich and Michalski (1983).

However, there is another way for hypothesis $i + 1$ not to be a proper subset of hypothesis i and so meet the Subset Principle. Hypotheses could partially overlap; in this case, that would mean that there is some example not covered by one guess that is in the second. Then the learner need not subscribe to minimal generalization, as discussed earlier. Angluin's theorem covers both sorts of cases.

Since the subset arrangement is the focus in the next two sections, it will be described in more detail. Suppose that all the possible target concepts or languages can be arranged in a nested order, like concentric circles. Then to meet the Subset Principle the acquisition procedure must order its hypotheses so that it always guesses the narrowest possible hypothesis or language at each step. This is because if all hypotheses are so nested then the only way for the hypothesis guessed at step $i + 1$ *not* to be a proper subset of that guessed at step i is for it to contain hypothesis i. The right sequence of guesses will be monotonically increasing—each description will cover the one before it in the sequence. This is what is meant by "timid acquisition"; at each step the system will take the smallest possible step consistent with evidence seen in order to avoid the possibility of guessing too large a language. It corresponds to an incremental search through the space of hypotheses, starting from the most specific first. The power of this principle suggests that we look for evidence that it is applied in natural learning settings. In the next two sections we shall see if we can find such evidence. As far as can be determined, the Subset Principle exhausts what can be said about ordering constraints in language acquisition and perhaps in other domains as well.

21.3 CONCEPT DEVELOPMENT AND THE SUBSET PRINCIPLE

The first example presented here is drawn from research on conceptual development initiated by Sommers (1971) and pursued by Keil (1979). It will be seen that children's developing knowledge about the world obeys the Subset Principle. But first let us summarize Keil's research. Keil claims that if one arranges a person's judgments of whether a set of "predications" of terms "makes sense" or not, then one obtains a characteristically hierarchical tree. By *predication* here Keil means such things as *is loved, is an hour long, can be thought about, is green,* and so forth. Thus, an apple can be green but not an hour long; a recess can be an hour long but not green; and both can be thought about. This is really just a way of describing how people categorize things. For instance, a recess and an apple are different things, because apples can be green but recesses cannot be. As Keil suggests, following Sommers, these facts can be represented by predicates placed at the nodes of a graph in such a way that interior nodes of the graph denote predicates and leaf nodes of the graph (those nodes not dominating any other nodes) denote things in the world like recesses and apples. A leaf is dominated by the predicate nodes that it makes sense to apply to that leaf. For example, since *recess* can be an hour long or can be thought about but cannot be green, it is placed below the first two nodes but not the last. But since an *apple* can be green or thought about but not an hour long, *apple* is placed below the nodes *can be thought about* and *is green.* Finally, predicates (the interior nodes of the graph) are placed according to the leaves they span. A node is dominated by another node if the leaves the first node dominates are a proper subset of the leaves the second node dominates. Applying this to the example produces the graph shown in figure 21-1. Objects in the world (tree leaves) are shown in italics.

Note that the predicate *is thought about* will typically be at the root of such a graph. More interestingly, the graph is almost always a tree. The key point is that one rarely finds a natural conceptual structure where the resulting graph forms M- or W-shaped patterns—that is, a case where a single term is subsumed by predicates from two separate hierarchical trees. This is because such a structure leads to indeterminate things in the world, as Keil notes. This is dubbed the *M-constraint.*

For example, suppose that a *zorch* was a word denoting either a blue pyramid or a red cube. Further suppose that *pyramids* could be blue or thought about and that *cubes* could be red or colored or thought about. The predication tree would then look like the one shown in figure 21-2.

According to the M-constraint, *zorch* could not stand for a natural concept, at least not in the vocabulary of blocks used earlier. This is because *zorch* would fall under two separate hierarchy trees, violating the M-constraint; note the distinctive partial W-shaped arrangement of the links of the graph at the bottom.

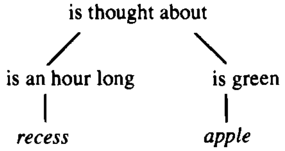

Figure 21-1: A simple predication tree.

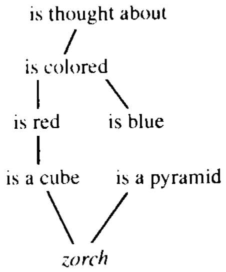

Figure 21-2: An unnatural concept results in an M-constraint violation.

Keil developed a method to describe the trees of developing children. Basically, he asked children whether recesses could be green, or if they could be an hour long, and so forth. He found that predication trees "grow" by the refinement of existing tree links without the destruction of existing domination relationships. No radical surgery occurs in which a pattern of domination links is completely destroyed. To see what this means, consider figures 21-3 and 21-4, which show a

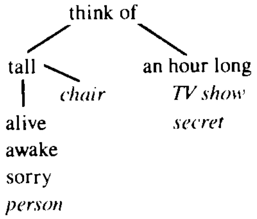

Figure 21-3: Sample predication tree at age five–six. (This is a single individual's tree.)

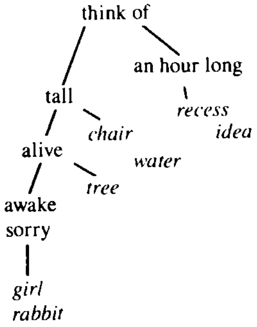

Figure 21-4: Foliated predication tree at age seven–eight.

sample of evidence that Keil obtained of children's predication trees at ages five to six and then at ages seven to eight.

At the earliest ages studied (five–six years), some children's predictability trees looked like the one in figure 21-3.

The trees of second graders were foliated versions of initial trees of this kind; that is, the new predication trees developed without existing domination links being destroyed; new links were simply inserted between existing predicates. Figure 21-5 illustrates this. Note how the class {alive, awake, sorry} is split into two. This is not a necessary condition for the development of predication trees; for example, trees could develop by a general rearrangement of predicate links. It could have been that children first consider *an hour long* to fall between *tall* and *alive*, only to move it from this position to the position shown in figure 21-4. But this evidently does not happen.

What Keil did not explain was *why* predication trees develop by branching. The Subset Principle can tell us why. Basically, this kind of branching corresponds to a "timid" refinement strategy, in the sense that there are no other refinements that could be interposed between the new tree and the old one. In other words, the children construct minimal extensions of their ways of categorizing objects in the world. Presumably the extensions are minimal in order to avoid overgeneralization. For example, a first-grade child could take the tree in figure 21-4 and make an "inductive leap" to a tree of the kind shown in figure 21-5 that splits apart *alive*, *awake*, and *sorry* all in one step.

If the child made this leap in one step, it might go astray: the correct tree could be one where *sorry* and *awake* were collapsed together. Note how this tree can be interposed between the current tree and the overly general guess. To avoid this possibility, children stick their ontological necks out as little as possible—at least, that is what is suggested by this evidence. Incremental refinement is not necessary, but apparently it is observed. The Subset Principle dictates what the next possible set of predication trees looks like: it should be some minimal refinement of existing trees.

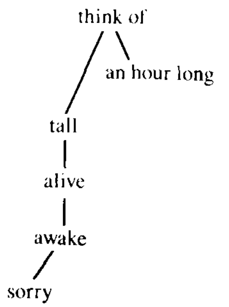

Figure 21-5: A predication tree that is possibly too general.

21.4 LEARNING SOUND SYSTEMS AND THE SUBSET PRINCIPLE

Let us now turn to the question of how the sound systems of language are acquired. Once again, some relevant background material and terminology will be presented first, and it will then be shown that the Subset Principle can actually help explain why only certain sound systems exist in the world's languages. This is an example of how a *general* learning principle may be used to account for a domain-specific constraint, in this case a constraint on what is a possible human language.

According to the distinctive feature theory of sound systems originally developed by Prague school structuralists such as Jakobson and pursued by Chomsky and Halle in *The Sound Pattern of English* (1968), all natural sounds such as *a* or *p* can be described via a small number of binary-valued distinctive features. By and large these features have an articulatory or an acoustic grounding, with names suggestive of how the tongue and lips are placed as they are pronounced, such as *high*, *back*, *anterior*, and the like. *Back* refers to a vowel sound produced in the back of the mouth; *anterior*, in the front of the mouth; *high*, with the tongue raised high. There are about twenty-four features in all. Given binary values for distinctive features (+ or −), there are 2^{24} possible single language sounds or *segments* (about sixteen million) and even more possible subsets of these segments, what are called *segmental systems*. However, most of these segmental systems are not attested in human languages. Why is this?

In part, the reason for this is that distinctive features are not determined independently of one another. Rather, certain distinctive features can be fixed *only after* certain other features are set. For instance, according to the theory of Kean (1974), the distinctive feature *consonantal* must be set before the feature *back* or *continuant*. (*Consonantal* is simply a binary feature that is + if a sound is consonant and − if the sound is not a consonant. *Continuant* is a sound produced like a continuous tone.)

Kean developed this theory as a way to explain some of the observed restrictions on possible sound systems and possible phonological rule systems. But there is another way to interpret such a theory, and that is as a developmental program for how a sound system is acquired. By construing the theory in this new way, one can in fact exhibit an acquisition system in which large numbers of developmental pathways are eliminated because of the *order* in which a small number of parameters are set. Kean's theory of markedness for phonological sounds (or segments) will be outlined here to show how this approach works in detail. Kean states the basic aim of her theory as follows:

> It is assumed here that there is a relatively small set of distinctive features with binary specifications in terms of which all the members of every segmental system can be characterized at every stage of phonological representation. The postulation of such a set of features makes a substantive claim as to the class of possible elements in phonological systems.
>
> Of the set of possible segments characterized by the distinctive features, it is evident that some are present in nearly every language, with others only occasionally occurring. For example, the segments *t* and *a* are nearly ubiquitous in segmental systems; they are

> found at all stages of phonological representation in an overwhelming majority of languages, but the segments *kp* and *u* only occasionally enjoy a place in segmental systems. The simple postulation of a set of features cannot account for such facts. (1974, 6)

To explain the relative frequency or rarity of certain sounds, Kean posits "a hierarchy of features which is derivable from the intrinsic ordering . . . of markedness conventions" (1974, 81). For example, vowels are usually − anterior, consonants are + anterior. It is therefore highly unusual, or *marked,* for a vowel to have the feature + *anterior.* But vowels also have the feature − *consonantal* and consonants the feature + *consonantal.* Therefore, the feature *anterior* is correlated with that of *consonantal;* in the usual or unmarked case the following rule applies:

unmarked anterior → + anterior (if we already have determined the feature + *consonantal* for the segment)

unmarked anterior → − anterior (if we already have determined − *consonantal*)

From the complement of this rule, we obtain the convention for determining what the value of *anterior* should be if it is marked:

marked anterior → − anterior (if we already have determined + *consonantal*)

unmarked anterior → + anterior (if we have determined − *consonantal*)

Determining whether a sound segment is marked for anterior or not logically demands that the feature *consonantal* be determined first. If the sound is − consonantal, then the sound will usually be − anterior (the unmarked case); if + consonantal, the sound will usually be + anterior. Pursuing this approach, Kean goes on to show that whether the feature *back* is unmarked (expected) or marked (unexpected) depends on the value of the distinctive feature *anterior.* One obtains the following hierarchy of distinctive features: *consonantal, anterior, back.* If this analysis is applied to all twenty-four distinctive features that Kean considers, one arrives at the dependency diagram shown in figure 21-6. This gives a complete picture of the *order* in which features must be set for the feature values for any sound segment like *p* or *t* to be determined.

Here is what the other feature names mean: *sonorant* is, literally, a sonorous sound; *low* is a sound produced with the tongue low; *labial,* with the lips; *lateral,* with the tongue at the side of the lips and mouth; *coronal,* midway up; *flap* and *trill,* by vibrating the tongue; *delayed release,* by blocking air and then exploding it outwards. The other feature names are mostly self-explanatory.

Each distinctive feature in the network depends on those features immediately *above* it to determine whether it is marked or not. For example, to determine whether the feature *continuant* is marked or unmarked we must know the values of the features *coronal* and *nasal;* to know whether *continuant* is marked or not, we must know the values of the features *coronal* and all features above *coronal, nasal,* and *sonorant.*

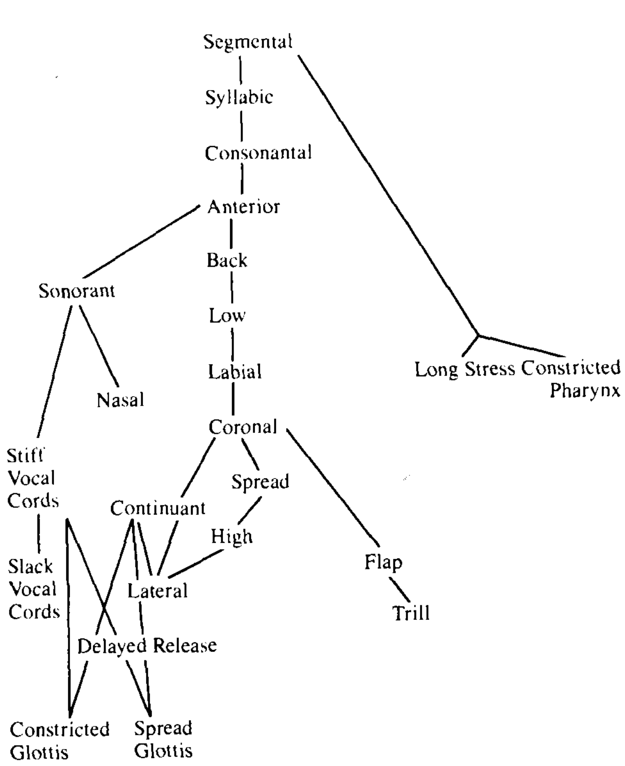

Figure 21-6: Hierarchy diagram for distinctive features.

Although Kean did not choose to do so, we may interpret this hierarchical structure as the specification of an acquisition procedure for learning a sound system. The key point is that this procedure follows the Subset Principle, and this explains the appearance of the dependency diagram. We now describe how this works.

According to distinctive feature theory, sounds can be distinguished only if they have different values for at least one of the twenty-four distinctive features. For example, the sounds *a* and *i* are distinguishable given a sound system that sets all distinctive features for the two sounds to the same value save for the feature *back.* The sound *a* is unmarked for back, but *i* is marked for back (i.e., is expected). As another example, the sound *ae* is also marked for back (− *back*) but is distinguishable from *i* because it is additionally marked + *low.*

We see then that a sound must be explicitly marked in order for it to be distinguished from the default set of plus and minus values. Otherwise, all sound would be unmarked for all distinctive features, and hence all would possess the same array of distinctive feature values. In other words, if the array of distinctive feature marks is regarded as partitioning the universe of possible sounds into equivalence classes, then if no sounds were marked there would be just one class of sounds, the totally unmarked one.

This remark is not quite accurate, however, since Kean also assumes a basic syllabic/nonsyllabic distinction in addition to purely distinctive feature contrasts. As a result, there is always an initial division of all possible sounds into two classes, consonants and vowels, according to the following rules:

unmarked consonantal → + consonantal (if segment is already − syllabic)
marked consonantal → − consonantal (if segment is already + syllabic)

Given the initial partition defined by the feature *syllabic*, we can thus distinguish two classes of sounds, even if no other distinctive features are used for marking sounds:

{*i, e, ae, u, o, oe, i, e, a u, o*} (+ syllabic)
{*p, t, t'*, . . .} (− syllabic)

At this stage then, there are in effect just two sounds, "consonants" and "vowels," as defined by local sound context. To generate new classes, we must mark additional distinctive feature values. The key idea here is that there is a definite order in which new features are used to form new sound classes. New classes may be formed by splitting established classes, with the split based on the order given by the distinctive feature hierarchy. Suppose we start with a division into just two classes of sounds, consonants and vowels. The next partition is based on the next *unused* (previously unmarked) feature in the hierarchy. This is a natural assumption. We cannot get a new class of sounds unless we explicitly mark a distinctive feature contrary to its expected value—otherwise, we would simply obtain the default feature settings for all later features in the feature hierarchy. So we must mark at least one new distinctive feature. Further, since features lower down in the hierarchy depend on the values of features above them, the natural place to look for the next distinctive feature to mark or not is the next feature below *consonantal* in the hierarchy, i.e., either the feature *anterior* or the feature *sonorant*. We cannot skip either of these features to try to mark, say, the feature *labial*, because the value of *labial* depends upon whether *low* and *labial* were marked or not, and these features have not yet been evaluated. So let us say that, in general, a new partition must be formed by marking exactly one of the distinctive features immediately below the last feature that was marked.

How is a split triggered? This choice must be "data driven" since different sound systems will have different sounds (from the adult point of view) that are marked for a particular distinctive feature. For example, as Kean observes, in Hawaiian only the sound *n* is marked for *sonorant*, but in Wichita, it is *r* that is so marked (1974, 57). A split must therefore be triggered by a detectable difference between at least one of the members of an existing sound class and the rest of the members of that class. Presumably, this difference could be detected on a variety of grounds, articulatory or acoustic. The new sound might just *sound* different. Nothing more will be said here about just how this might occur. What one can say, however, is just *where* the next distinction will be made. The Subset Principle states that the next available unused distinctive feature in the Kean hierarchy must be used as the point of refinement. Otherwise the learner could guess too large a language and go astray.

As an example, consider again the class of vowels {*i, e, ae, eu*, . . .}. According to the feature hierarchy diagram, the next split of this class must be described by the value of the next feature below *consonantal*, namely, *anterior*. As it turns out, the feature combination [− *consonantal* + *anterior*] is physically impossible (the mouth and tongue cannot produce both features simultaneously) so that in fact the feature *anterior* cannot be freely varied given that the value of the feature *consonantal* is minus. So the candidate distinctive features that may be used to split the class {*i, e, a*, . . .} become the features just below *anterior*, namely, *back* or *sonorant*. The combination [− *consonantal*, − *sonorant*] is also impossible, however, so that a potential split must be pursued by considering *back*. Features such as *strident* or *continuant* would not be used as this point.

Suppose then that the feature *back* is selected for marking, forming the basis for a new partition of sounds. By marking *back* we obtain the following potential classes: *marked back* {*i. e, ae, u*, . . .} and *unmarked back* {*a*, etc.}. Kean's marking convention *unmarked back* → + *back* given − *anterior* establishes that *marked back* must be − *back* in this case, and *unmarked back*, + *back*. In effect, two kinds of "vowels" have been established, corresponding to two possible pathways through the hierarchy diagram.

The important feature of the partitioning process is that splitting occurs at the leading edge of the directed hierarchy graph by successive refinement of exiting classes of sounds. This is a powerful constraint on possible natural sound systems. Suppose that this constraint did not exist. Then it would be possible to have a sound system in which the feature *sonorant* was not used—not set as either marked or unmarked—but the feature *labial* was used. There would be a "gap" in the feature hierarchy skipping over the use of a feature. No such system exists among the world's languages.

Because extension of classes occurs solely via the refinement of existing partitions, the set of sound classes at step *i* will be a refinement of all of those before it in the developmental sequence. This is simply the same constraint we saw with the predication trees, now repeated in a quite different domain. In particular, this restriction means that just *one* distinctive feature will be set as either marked or unmarked at any single acquisition step. Again, this is not a necessary constraint, since it is not clear why one could not develop a new class by marking two or more features in one step.

The effect of this constraint is to guarantee incremental acquisition. At any step *i* in the development of a sound system, the classes of sounds will be at most one mark (*m*) different. For instance, this constraint excludes the array of marks described in figure 21-7, where *m* stands for a marked feature and *u* for an unmarked feature.

	X1	X2	X3	X4
Consonantal	*m*	*u*	*u*	*u*
Anterior	*m*	*u*	*u*	*u*
Back	*m*	*u*	*u*	*u*
Low	*u*	*m*	*u*	*u*
Labial	*u*	*u*	*m*	*u*
Sonorant	*u*	*u*	*m*	*u*

Figure 21-7: An impossible configuration of *u* and *m* marks.

From one point of view the one-mark constraint is a puzzling one. It is not at all obvious why sound systems should be designed so that the alteration of a single distinctive feature could convert an *a* into an *i*. This would seem to be an unwise design choice from the standpoint of error detection or error correction; as is well known, in order to be able to correct errors of *k* bits, then sounds would have to be separated by a ball of radius $2k + 1$ (since one must guarantee that changes of up to *k* bits in any two sounds still leave one able to determine the original sound).

Importantly, natural sound systems do seem to obey the one-mark constraint, as Kean observes. In other words, the matrices of *m*'s and *u*'s of natural sound systems cannot look like the one depicted in figure 21-7, with no sound more than one *m* away from any other. That this is so may be attributed to a design that follows the Subset Principle. Let us see why this is so.

As the way in which the hierarchy can be interpreted as an acquisition model has been described here, only one feature can be used to form a new partition of sounds—only one *mark* (*m*) is ever added at any given step. As a result, at any stage in the acquisition of a sound system the partitions correspond to sounds that are at most one *m* apart, automatically satisfying the distinguishability constraint. So Kean's observation might well be explained as a side effect of the acquisition of sound systems. Even so, it seems as though a stipulation about the well-formedness of segmental systems has merely been replaced with a stipulation about the acquisition of segmental systems. Why should acquisition be incremental?

Suppose that acquisition is not incremental and that two or more marks can be added at a single step. It would then be possible to form a new class partition based on marking both the features *labial* and *sonorant* without having first used the feature *sonorant* to form any sound classes.

There would be no class that would correctly accommodate a sound that is labeled [*unmarked labial, marked sonorant*]. One way to remedy this problem would be to allow the procedure to go back and rebuild classes that have already been formed, but this would violate the developmental ordering that has been assumed. The fringe of the hierarchy tree would no longer summarize the possible next states that could be hypothesized, since there could be sounds such as *n* that would demand the interpolation of new classes between older partitions and the current partition. In other words, the one-*m* constraint amounts to the demand that new classes be the minimally specific refinements of existing classes. It is impossible to guess an overly general sound system, because each new guess is the smallest possible refinement of preceding guesses. But this is just the Subset Principle again. At each step, the narrowest language is hypothesized, consistent with positive evidence seen so far.

21.5 LEARNING SYNTAX AND THE SUBSET PRINCIPLE

The Subset Principle subsumes a variety of proposals that have been advanced in the linguistic literature that order hypotheses for language acquisition. In fact, it appears as though the Subset Principle exhausts what can be said about ordering constraints in acquisition. In support of this claim several proposals that have been made regarding the ordering of hypotheses in the acquisition of syntactic constructions will be reviewed. It is not important that the reader appreciate all the details of these examples. The intent is to give a feel for the variety of different kinds of grammatical constructions that fall under the Subset Principle.

21.5.1 An Adjacency Requirement in English

In English, noun phrase direct objects must be adjacent to verbs: *I gave a book quickly to Bill* is fine, but *I gave quickly a book to Bill* is not. (In some languages, such as French, this constraint is weakened so that an adverb may be interpolated between verb and object; in other languages, such as Japanese, this constraint is so weak that the object can be quite distant from the verb.)

How is the adjacency requirement acquired? Once again, the Subset Principle may be invoked. The most restrictive assumption possible is that adjacency holds since it generates the *narrowest* class of language possibilities. To assume otherwise would be to guess a language that could be too large, hence a possible Subset violation. A language satisfying the adjacency condition could be a proper subset of one that was not and yet cover the same triggering data. The acquisition procedure thus assumes an adjacency requirement as the default, unmarked case, loosening it only if positive examples are encountered that indicate violations of adjacency. Since examples violating adjacency (*I hit hardly Bill*) will never be encountered in English, this strict requirement will never be dropped.

In other languages (like French) positive examples exhibiting adjacency violations would prompt a relaxation of these conditions, perhaps along a continuum of possibilities. Thus one might expect to find languages where strict adjacency was relaxed according to a hierarchy of phrasal types. This prediction seems to be confirmed.

21.5.2 Arguments of Verbs

Verbs differ in the number of noun phrase objects (or arguments) that they require and in whether those arguments are obligatory or optional. For example, *eat* may or may not take an argument denoting the thing eaten: *John ate an ice cream cone, John ate.* In contrast, *take* must take an argument: *John took an ice cream cone* is fine, but *John took* is not. Note that a language where a verb may or may not take an argument is a superset of a language where that verb must take an argument. If hypotheses are to be ordered by the Subset Principle, the first guess to make about any verb is that if it appears with an argument, then that argument must be assumed obligatory until a positive example appears in which that argument is not present at all; if such an example appears, the argument is optional. This strategy is observed in children (Roeper, 1982).

21.5.3 Bounding Nodes for Subjacency (Rizzi, 1978)

In most current theories of generative grammar, it is assumed that grammatical rules obey a certain "locality principle," in that a movement cannot cross more than a single sentence boundary. For example, in the first sentence below, *John* is understood as the subject of the embedded sentence *to like ice cream.* The second sentence is ill formed if interpreted this way. The only difference is that the second sentence interposes an additional sentence boundary via the *it is certain* clause. Square brackets mark these boundaries.

1. John is certain [$_S$ trace to like ice cream]
2. John seems {$_S$ it is certain [$_S$ trace to like ice cream]]

This called the subjacency constraint. This constraint is also what makes the following sentence poor, where the *who* words are linked to positions as indicated by subscripts *i* and *j*. For example, the first *who* is the object of *know.* Unfortunately, two S's must be crossed to link up to this position—hence the sentence is no good.

The man who I don't know who knows.

The man [who$_i$[$_S$ (first S) I don't know [who$_j$ [$_S$ (second S) *j* knows *i*]]]]

Interestingly, this last sentence is grammatical in Italian, as discussed by Rizzi (1978):

L'uomo {*wh*$_i$ che non so [chi$_j$ {$_S$ *j* conosca *i*]]]}

According to Rizzi, this is because it is a *full* clause with *that* or *for* in it that counts for subjacency in Italian, not just a simple "S" or sentence. (A full "S" in English would be something like *For John to go* . . . .) This kind of phrase is called an S-bar. Rizzi claims that S-bar, not S, is what counts in Italian. Therefore, the *who* in the sentence above can be the object of *conosca* (know) because it crosses only a single full clause boundary that starts at *chi.* The second boundary is an S, not an S-bar. Apparently, the choice of a bounding node is yet another parameter that must be set in order to learn a language.

Suppose Rizzi's analysis is correct. How could the choice of bounding node be determined on the basis of evidence received by an acquisition procedure? Once again, let us apply the Subset Principle. If the bounding node for subjacency is S, then a narrower class of languages is generated than if the bounding node for subjacency is S-bar. Therefore, by the Subset Principle, the acquisition procedure's first hypothesis should be to set the bounding node for subjacency to S. In other words, the default assumption is that all languages are like English in this regard. If this assumption is wrong, then a positive example will appear that violates S-bounding—as in the Italian example above. Then the acquisition procedure can reset the subjacency parameter to the next "largest" value, namely, S-bar.

21.6 SUMMARY AND CONCLUSIONS

This chapter has shown that there is at least one quite general principle of learning, the Subset Principle, that applies "across the board" in a domain-independent fashion. The Subset Principle arranges the order of hypotheses that a learner should advance in the face of positive-only evidence. The principle has wide applications, showing up in such diverse domains as the acquisition of category concepts, sound systems, and syntax. It has been used explicitly in at least one model for the acquisition of language (Berwick, 1982) and implicitly in the version space model of acquisition.

The Subset Principle makes strong predictions about the order of events in human language acquisition. Are any of these predictions confirmed? In fact, recent experimental tests have been made of the ordering constraints implied by the acquisition proposals described in this section. Children are asked to "act out" certain situations with toy animals in order to see if they understand particular sentences of the sort described above. The results are preliminary (Wexler, 1984), but so far correspond exactly to the predictions of the Subset Principle.

It has already been seen that Keil's work in concept acquisition points to confirmation of the Subset Principle. What of the acquisition of language sound systems? Results here are sketchy. However, there is at least one "classic" piece of evidence, namely, the observations of Jakobson (1968). The Subset Principle ordering predicts that *t, p,* and *k* would be among the first consonants acquired and *a, i,* the first vowels. This sequencing appears to be *roughly* verified by empirical work, though there has been controversy regarding Jakobson's more restricted and probably overly strong proposal.

It remains to be seen whether other kinds of human learning, such as the acquisition of arithmetic skills, abide by the Subset Principle. There is at least some suggestive evidence (see VanLehn, 1983) that they do. Several natural learning systems, then, obey the Subset Principle, in which positive-only example evidence plays a dominant role. Machine learning systems would do well to follow this successful design.

ACKNOWLEDGMENTS

This chapter describes research carried out at the Artificial Intelligence Laboratory at MIT. Support for the Laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency under Office of Naval Research Contract N00014-80-C-0505.

References

Angluin, D., "Inductive Inference of Formal Languages from Positive Data," *Information and Control*, Vol. 45, pp. 117-35, 1978.

Berwick, R., "Computational Analogs of Constraints on Grammars," *Proceedings of the Eighteenth Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pa., pp. 49-54, 1980.

———, "Locality Principles and the Acquisition of Syntactic Knowledge," Ph.D. diss., Department of Electrical Engineering and Computer Science, MIT, 1982.

Brown, R., and Hanlon, C., "Derivational Complexity and the Order of Acquisition in Child Speech," in *Cognition and the Development of Language*, J. R. Hayes (Ed.), Wiley, New York, 1970.

Chomsky, N., *Rules and Representations*, New York, Columbia University Press, 1980.

Chomsky, N., and Halle, M., *The Sound Pattern of English*, New York, Harper and Row, 1968.

Dietterich, T., and Michalski, R., "*A Comparative Review of Selected Methods for Learning from Examples*," in *Machine Learning: An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (Eds.), Tioga, Palo Alto, Calif., 1983.

Gold, E., "Language Identification to the Limit," *Information and Control*, Vol. 10, pp. 447-74, 1967.

Jakobson, R., *Child Language, Aphasia, and Phonological Universals*, The Hague, Mouton, 1968.

Kean, M., "The Theory of Markedness in Generative Grammar," Ph.D. diss. Department of Linguistics, MIT, 1974.

Keil, F., *Semantic and Conceptual Development: An Ontological Perspective*, Harvard University Press, Cambridge, 1979.

Marshall, J., "Language Acquisition in a Biological Framework," in *Language Acquisition*, P. Fletcher, and M. Garman (Eds.), Cambridge University Press, New York, 1979.

Mitchell, T., "Version Spaces: An Approach to Concept Learning," Computer Science Report CS-78-711, Stanford University, 1978.

Rizzi, L., "A Restructuring Rule in Italian Syntax," in *Transformational Studies in European Languages*, S. J. Keyser (Ed.), MIT Press, Cambridge, 1978.

Roeper, T., "On the Deductive Model and the Role of Productive Morphology," in *The Logical Problem of Language Acquisition*, C. Baker and J. McCarthy (Eds.), MIT Press, Cambridge, 1982.

Sommers, F., "Structural Ontology," *Philosophia*, Vol. 1, pp. 79-85, 1971.

VanLehn, K., "Validating a Model of Children's Arithmetic Skills: Sierra," *Proceedings of the International Machine Learning Workshop*, R. S. Michalski (Ed.), Allerton House, University of Illinois at Urbana-Champaign, June 22-24, 1983.

Wexler, K., "Independence and the Subset Principle," University of Massachusetts *Conference on Formal Models of Language Acquisiton*, Amherst, 1984, forthcoming.

Wexler, K., and Culicover, P., *Formal Principles of Language Acquisition*, MIT Press, Cambridge, 1980.

Winston, P., "Learning Structural Descriptions of Blocks World Scenes from Examples," in *The Psychology of Computer Vision*, P. H. Winston (Ed.), McGraw-Hill, New York, 1975.