Chapter 4

Dynamically guided learning

Rebecca Gómez

Abstract
Recent research on human learning has revealed a pervasive ability to track statistical structure in adulthood and infancy. Because statistical information abounds in visual and linguistic structure, statistical learning has potential for playing an important role in the acquisition of complex skill. This chapter summarizes the literature on statistical learning and evaluates it in terms of its potential for circumventing problems that have been traditionally posed for learners, such as over generalization and how learners choose among multiple forms of structure. Empirical findings suggest that statistical learning is highly data driven. Additionally, learning appears to be constrained by preferred structure (structure that is particularly salient or easy to process), but also by the pressures exerted by the environment (learners tend to seek out invariant structure). Learning arises in the interaction of these two forms of constraint, resulting in a dynamically guided process.

4.1 Introduction
Statistical structure abounds in perceptual information. Whether auditory, visual, or tactile, stimulation is rich in frequent associations and patterns. Humans show a striking capacity for tracking such information. ‘Statistical’ learning, as this capacity has come to be known, is at the very root of development as evidenced by its emergence in early infancy (Gómez and Gerken 1999; Kirkham et al. 2002; Saffran et al. 1996). Research in this area complements the two plus decades of work on computational models of learning and has been a major contributor to a renaissance of learning in psychological research.

For many years, especially in the area of language acquisition, learning was assumed to have little influence on language development other than explaining how children acquire the variations between different languages (see Chomsky 1965). Rather, innate
mechanisms were thought, almost solely, responsible for guiding acquisition. Such mechanisms were not only language specific, but were thought to operate at a fairly high level (for instance the knowledge that sentences in one's language either do or do not require overt subjects). However, recent findings raise key questions about the form of innate sensitivities and the interplay of these with learning. Although innate sensitivities are most certainly involved in development, characterizing them accurately may be dependent on a more thorough understanding of learning. Identifying the scope and limits of learning should bring us closer to specifying the initial state.

I have organized this paper with two goals in mind. One is to review key literature on statistical learning and to convey an up-to-date summary of this work. I will do this with an eye toward evaluating how this research addresses fundamental challenges to learning theories having to do with understanding how learning might be constrained, how learners might recover from faulty generalizations, and how they might choose among multiple kinds of structure. The second goal is to explore the extent to which current work on learning addresses these challenges. Much work on statistical learning has been conducted in the context of language acquisition, and thus this chapter will reflect that perspective, but the issues generalize to other developmental domains.

The organization of the paper is as follows. First I will outline two fundamental issues for learning theories—how learners recover from overgeneralizations and the question of how learners choose among multiple kinds of structure. Next I will summarize the literature on statistical learning. I will go on to explore a mechanism for understanding how learners might choose among multiple kinds of structure and, in doing so, will consider the extent to which learners are likely to overgeneralize. Finally, I will highlight recent work in the area of memory and its implications for theories of learning.

4.1.1 Fundamental learning issues

There are numerous puzzles and questions that can be raised about basic learning mechanisms. Are they domain general or domain specific? Are they fundamentally different for child and adult learners? How rapid is learning? How permanent is it? What is the connection between learning and memory? What types of learning mechanisms do humans share with other species? What is the relationship between innate abilities and learning? Where do innate sensitivities leave off and where does learning begin? Obtaining answers to these questions is challenging, but not impossible. There are problems of a logical nature, however, that seem insurmountable – in particular, the problem of how learners recover from erroneous generalizations.

Pinker (1995) outlines four ways in which children's hypothesizes may differ from the rules of the target language and thus how children might acquire erroneous information. First, children's hypotheses may be entirely non-overlapping with the target language. This would be the case if children had only ungrammatical sentences in their repertoire. The second is that some, but not all, of children's hypotheses
overlap with the target language. In this case, they would use some grammatical sentences but would also use ungrammatical ones. The third scenario is that children’s hypotheses form a subset of the target language such that all of their sentences are grammatical, but they have not yet mastered all of the grammatical structures of the target language. None of these scenarios pose major problems for children because they will presumably continue to generate hypotheses in the course of continued (or ‘positive’) exposure. However, the fourth scenario, in which children’s hypothesized language is a superset of the target language, does pose a problem, because the only way children can recover from such errors is by exposure to negative evidence (Gold 1967). This is because positive evidence only continues to confirm the overly general hypotheses. Negative evidence could come in the form of parental feedback; however such feedback appears to occur only sporadically (Marcus 1993) and, when present, has minimal impact (Brown and Hanlon 1970). As noted in Pinker (1995), without negative evidence children ‘must have some mechanism that either avoids generating too large a language—the child would be conservative— or that can recover from such overgeneration’ (p. 153). Otherwise, children will not be able to converge on their target language.

This puzzle has figured prominently in the literature on language learnability (e.g. Gold 1967; Wexler and Culicover 1989) and, in turn, has greatly influenced fundamental assumptions in language acquisition theory. Yet work on this problem has primarily been conducted using formal mathematical approaches and thus has tended to ignore key aspects of human cognition. Taking human cognition into account might lead us to question these assumptions. For instance, we might ask whether humans are actually likely to overgeneralize. We might also inquire about the nature of human memory to determine the extent to which overgeneralizations are likely to survive in memory. Answers to such questions may shed a different light on the degree to which learning may be incapacitated by the generation of erroneous hypotheses.

A second related challenge is the question of how learners choose among multiple competing hypotheses. Traditional views of learning have assumed that learning mechanisms would be too rudimentary to separate relevant from irrelevant structure and instead would make all possible generalizations, leading to a combinatorial explosion of possibilities (Pinker 1984). As with the problem of overgeneralization, this problem has been framed in formal terms instead of taking human cognitive capabilities into account. However, it is important to ask whether the problem of combinatorial explosions in human learning is a realistic concern. To determine this, we must characterize human learning mechanisms more fully. I will explore these issues in detail in ensuing sections.

4.1.2 A few notes about methodology
First, I will convey some key methodological details. Infants learn a great deal about natural language in a few short months, so in order to assess their learning abilities,
they must be exposed to novel stimuli. Additionally, natural language is rich in correlated cues, making it difficult to pinpoint the exact source of learning. Artificial language materials are thus used in statistical learning studies to control for prior learning and to have more precise control over the types of cues presented to learners. Because the materials are novel they can also be used to assess learning in adults.

Statistical learning procedures usually involve two phases, a familiarization phase followed by a test. Length of familiarization varies from 2 to 3 min in infant studies to 10 to 20 min (or more) in adult studies. Adult participants are often used to obtain detailed information that cannot easily be obtained from infants and to investigate age-independent learning. Most studies use a two-language design so that half of the learners are exposed to Language A and half to Language B. At test, learners are exposed to strings from both languages. Language A strings violate the constraints of Language B and vice versa. Thus, both groups receive the same test but what is grammatical for one group is ungrammatical for the other. This design ensures that the structure of the languages, as opposed to something idiosyncratic about one language or the other, is responsible for learning. Great care is taken in the design of the materials to ensure that the critical grammatical structure will drive learning.

Infants are tested using visual or auditory preference procedures that record the amount of time they attend to different stimulus types. Adult learning is assessed by means of grammaticality judgments or two-alternative forced choice. In all cases, learners discriminate between strings that conform to their training language and strings that do not. Adult learning gains tend to be modest, hovering between 60 and 70 per cent.

4.2 Statistical learning

Statistical learning is the discovery of statistical structure in the environment. Statistical structure can take many forms, including the frequency of cues or events, the co-occurrence frequency of cues or events, or the conditional probability on one cue or event given another. As an example, co-occurrence frequency is defined as the probability of two events occurring together. Conditional probability is the probability of the occurrence of an event given that another event has previously occurred; the probability of Event B given Event A is the probability of Event A and Event B divided by the probability of Event A, \( P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \). There are other forms of statistical structure but common to all forms is the characteristic that cues or events occur with some regularity.

Much has been made of the fact that in some studies infants listen longer to strings from their training language whereas in others they listen longer to the other language. Although there is some reason to think that direction of preference may be linked to the complexity of the stimulus (Hunter and Ames 1988), this issue is largely unresolved. Because it is impossible to predict the direction of preference with absolute certainty, orienting reliably longer to one stimulus type or the other is taken as an indication of discrimination, regardless of direction.
Sensitivity to statistical structure has been shown in infants as young as 2 months of age. Kirkham et al. (2002) habituated infants to a continuous stream of three randomly ordered shape pairs (e.g. pink diamond, green triangle, yellow circle, turquoise square, blue cross, red octagon). Shapes followed each other with 1.0 probability within pairs and with a 0.33 probability between pairs. Additionally, all shapes were equally frequent. Because the shapes were presented continuously, looking time differences to legal versus illegal combinations at test would indicate sensitivity to the statistical information. Infants looked longer on trials containing illegal shape sequences suggesting that they had tracked the statistical information. Because sequence pairs occurred more frequently than non-sequence pairs, it is difficult to know whether infants of this age are tracking conditional probabilities between adjacent events, the co-occurrence frequency of events, or both.

However, infants are definitely able to track conditional probabilities by 7 to 8 months of age (Aslin et al. 1998; Fiser and Aslin 2002a; Saffran et al. 1996). Saffran et al. (1996) exposed 8-month-olds to continuous streams of randomly ordered three-syllable words (e.g. tupiro, golatu). Additionally, the frequency of individual syllables was held constant. The primary cue to word boundaries was the higher conditional probabilities occurring between syllables within words versus the lower conditional probabilities occurring for syllables spanning words. Infants were able to use this information to identify word boundaries in running speech. Although the co-occurrence frequency of syllables within words and syllables spanning word boundaries was not held constant in the initial study, Aslin et al. (1998) equated co-occurrence frequency during training and still showed learning. All studies reported by Saffran and colleagues since then have controlled for this variable. For instance, Thiessen and Saffran (2003) have reported learning of conditional probabilities by 7 months of age. Other studies have demonstrated similar learning with multielement scenes (Fiser and Aslin 2002a) and tone sequences (Saffran et al. 1999), and with non-human primates (Hauser et al. 2001), showing that this learning is not specific to language or to humans.

By 12 months, infants are able to track conditional probabilities of words in strings (Gómez and Gerken 1999; Saffran and Wilson 2003), and can generalize their knowledge of frequent patterns to strings in new vocabulary (Gómez and Gerken 1999; Marcus et al. 1999). Between 12 and 17 months of age infants have begun to abstract categories from words in sequence and can learn phrase structure relationships between them (Gómez and Lakusta 2004; Gerken et al. in press). By 15 months, infants are able to track non-adjacent dependencies in sequential structure (Gómez 2002; Gómez and Maye in press; Santelmann and Jusczyk 1998).

Infants also appear able to learn under especially challenging conditions. Saffran and Wilson (2003) investigated infants’ ability to use one source of information as input to another. Twelve-month-olds were familiarized with strings of novel words in continuous speech. Word ordering within strings was constrained by a finite-state
grammar such that only certain orders of words were legal. This stimulus design required infants to identify words before they could learn the relationships among them in sequence. Infants were able to discriminate legal from illegal strings at test, demonstrating that they had both segmented units at the word level and had learned how words could be combined to form sentences. In another study, Gómez and Lakusta (2004) asked whether infants would generalize even when inconsistent strings were present. They did this by manipulating the extent to which training strings were drawn from a predominant artificial language. Infants in an 84/16 condition heard strings from their predominant language 84 per cent of the time and strings from another counterbalanced language 16 per cent of the time (other infants were exposed to 100/0 or 68/32 ratios). Infants in the 84/16 condition performed as well as those in the 100/0 condition, showing that they were able to tolerate some inconsistency in their input. However, with a high percentage of intrusions (the 68/32 condition), infants failed to discriminate, demonstrating reasonable limits on this learning.

Finally, a number of reports show remarkable similarities in adult, infant, and child learning, suggesting that abilities are present early on and persist into adulthood. For instance, adults, children, and infants process conditional probabilities similarly such that they are able to detect words and tones in continuous speech (Saffran 2002; Saffran et al. 1996; Saffran et al. 1997; Saffran et al. 1999). They are equally adept at detecting conditional probabilities in sequentially presented visual stimuli (Fiser and Aslin 2002a, b). Infants and adults also appear to generalize similarly, regardless of whether they are abstracting patterns of repeating elements (Gómez and Gerken 1999; Gómez et al. 2000; Marcus et al. 1999) or category structure (Braine 1987; Gerken et al. in press; Gómez and Lakusta 2004). Finally, similar factors are at play in how infants and adults come to detect remote dependencies in sequential structure (Gómez 2002; Gómez and Maye in press).

This section paints a picture of significant learning capacities, with infants showing discrimination after brief exposure to fairly complex structure (e.g. Gerken et al. in press; Gómez 2002; Gómez and Maye in press; Gómez and Lakusta 2004; Saffran and Wilson 2003). But with such seeming raw power, what might constrain learning? This issue is important because the very success of statistical approaches hinges on learners’ ability to distinguish relevant from irrelevant information (Chomsky 1959; Pinker 1995). With such precocious abilities, what is to keep learners from making erroneous generalizations? In addition to the principles and parameters approach in which Universal Grammar is thought to consist of the principles common to all languages and the parameters that vary among languages (Chomsky 1981), several possibilities have been proposed. One is that humans are constrained by their perceptual apparatus to consider only certain possibilities (Saffran 2002, 2003; Newport and Aslin 2004). Another is that learning is constrained by the presence of correlated cues such that learning is more likely to be successful when multiple cues are present (e.g. Morgan and Demuth 1996; Morgan et al. 1987). A final constraint is prior learning,
the extent to which prior experience affects the type of structure learners seek out (Lany et al. 2004; Lany and Gómez 2004). We will go into these proposals in more detail in the next section.

4.3 Constraints on learning

The problem of how learning might be constrained is a particularly thorny one for statistical learning approaches. Criticisms are that learning is not sophisticated enough to cope with the complex structure found in language (Chomsky 1965), or that it is too powerful, resulting in attention to irrelevant information and unchecked overgeneralization (Pinker 1984, 1995). In response, a number of constraints have been proposed to explain how learners converge on appropriate structure.

A widely held view is that innate tendencies (congruent with linguistic universals) constrain language (Chomsky 1981). Such universals take the form of principles true of all languages and parameters that differ among languages. Parameters that vary among a small number of options are thought set on initial exposure to a particular language. However, there is little evidence that parameters are set in this manner. Rather, acquisition appears to occur over an extended period of time beginning early in infancy. Although the one-exposure criterion could be relaxed, this begins to move parameter setting into the realm of learning. Additionally, as outlined above, there is mounting evidence that statistical processes contribute to acquisition, suggesting that learning may be more central to language acquisition than previously thought. Saffran (2003) and Newport and Aslin (2004) have suggested that learning is constrained by human perceptual sensitivities that map onto the kinds of structure that occur across languages. Saffran (2003) refers to this as constrained statistical learning. Thus, instead of innate principles driving learning, human perceptual constraints guide what is learned such that some kinds of information are learned more readily than others, particularly in certain domains (e.g. Saffran 2002). Consistent with this proposal is the idea that languages are shaped by human information processing capabilities. Aspects of language that are particularly learnable are likely to be those that persist in the evolution of language. Thus, there is a symbiosis between language and human learning such that perceptual sensitivities are tuned to the structure of human language and in turn languages are constrained to be learnable by human perceptual systems (Bever 1970).

According to a multiple cue integration view, learning is constrained by multiple, overlapping cues that may be statistical, phonological, or prosodic in nature (Billman 1989; Christiansen and Dale 2001; Christiansen and Curtin 2004; Morgan and Demuth 1996; Morgan et al. 1987). Whereas no one cue may occur with enough predictability to support learning (Fernald and McRoberts 1996), multiple occurring cues are more reliable. In human language learning, multiple cue integration results in better generalization (Billman 1989; Braine 1987; Frigo and McDonald 1998; Gerken
et al. 1999; Gerken et al. in press). In connectionist models, multiple cue integration leads to faster, more uniform acquisition of syntax (Christiansen and Dale 2001), and to superior segmentation (Christiansen and Curtin 2004).

With respect to human learning, a case in point comes from work on generalization in which grammatical classes are given arbitrary labels such as a, X, b, and Y (Braine 1987; Frigo and McDonald 1998; Smith 1969). Words from these classes are combined to form legal phrases. For instance, aX and bY might be legal whereas aY and bX are not (this is analogous to relationships in English between determiners and nouns or auxiliaries and verbs, where children have to learn that determiners precede nouns but not verbs). Learners are exposed to most, but not all, aX and bY phrases, then are tested to see if they will discriminate legal phrases they have not yet encountered from illegal ones. As in natural language, ‘functor-like’ a- and b-categories have fewer members than ‘lexical-like’ Xs and Ys. Although learners readily acquire the legal positions of words in terms of which occur first versus second (Smith 1969), categories and their relationships (that words belong to particular a, b, X, and Y classes and that a-words go with Xs and not Ys) are virtually impossible to acquire unless some subset of the X- and Y-category members are marked with multiple conceptual or perceptual cues (Braine 1987; Frigo and McDonald 1998; Gerken et al. in press; Gómez and Lakusta 2004; Wilson 2002).

In the case of infants, Gerken et al. (in press) created a set of stimuli in which feminine lexical stems appeared with the case endings –oj and –u and masculine stems appeared with the case endings -ya and -em. Case endings in these experiments were equivalent to a- and b-elements. Additionally, cues distinguishing Xs and Ys were present for a subset of category members. For instance, some of the X-words contained the derivational suffix –k (e.g. ‘polkoj,’ ‘polku’) whereas some of the Y-words contained the suffix –tel (e.g. ‘zhitelja,’ ‘zhitelyem’). Seventeen-month-olds were first familiarized and then were tested to see if they would attend differentially to novel aX and bY stimuli versus ungrammatical aY and bX ones. Infants were able to generalize to new grammatical test items in which the suffixes were absent, generalizing to words such as ‘vannoj’ and ‘pisaren’ after hearing ‘vannu’ and ‘pisarya.’ As with adults, learning is impossible for infants to attain without multiple cues to category structure. Gómez and Lakusta (2004) investigated a precursor to this learning in 12-month-old infants.

Another type of constraint is prior experience. The importance of this constraint stems from the fact that, while key structure occurs in many forms in language not all forms occur with equal frequency or are as richly cued as others. For instance, nouns and verbs in English are both cued by co-occurring function morphemes (‘the,’ ‘a’ and ‘-s’ in the case of nouns and ‘was,’ ‘is,’ and ‘-ing’ in the case of verbs), but cues occur with greater statistical likelihood for nouns. Thus, children might be more attuned to the statistical information cueing noun categories than the information cueing verbs, and fail to learn the less well-cued form. However, learners can circumvent this
problem if they can use knowledge of a well-cued form to detect one that is less well cued. Consistent with this idea, Saffran and Theissen (2003) found that prior exposure to a particular phonological pattern influences which words infants are likely to identify in a segmentation task. Infants exposed to CVCV words in a prior training phase (C = consonant, V = vowel) listened longer to new CVCV words than to words of the form CVCCVC (the opposite held true for infants exposed to CVCCVC forms during training). In the realm of syntax acquisition, Lany et al. (2004; see also Lany and Gómez 2004) found that prior experience with syntactic categories can guide learners’ subsequent generalizations. In particular, after exposure to a language involving a key syntactic relationship (the aX/bY language detailed above), learners were able to detect relationships between a-, X-, b-, and Y-elements in a more complex language involving acX and bcY structure. This language was particularly challenging because the intervening c-element required learners to keep track of non-adjacent dependencies, which are harder to acquire than adjacent ones (Newport and Aslin 2004). Additionally, the more complex language was instantiated in novel vocabulary, forcing learners to draw on their knowledge of the abstract structural relationships of the aX/bY language. In the absence of prior exposure to the more simple language, learners were unable to achieve the more complex generalization. This finding is important for showing how learners might build more complex structures by scaffolding on simpler ones.

Thus, constraints appear to play an important role in learning. In most of these studies researchers have asked whether learners can acquire some particular type of structure (frequency, co-occurrence, or conditional probabilities, etc.; see Gómez and Gerken 2000, 2001 for review), but in the real world children need to discriminate among multiple sources of structure. The next section addresses this topic.

4.4 Choosing among multiple kinds of structure

Learners faced with multiple types of structure have to determine which is relevant to the particular learning problem at hand. Various solutions to this problem weight the influence of innate sensitivities and sensitivity to statistical structure, and their degree of interaction, differentially. For instance, the emphasis in the principles and parameters view is predominantly on innate linguistic constraints. Structure in language plays only a minor role in triggering parameters. While the constrained statistical learning view emphasizes perceptual constraints on learning it does not specify the extent to which constraints interact with statistics when learners are faced with two or more kinds of information. Multiple cue integration emphasizes statistical information but does not specify how such information might interact with perceptual constraints. The prior experience view is agnostic on this issue. Thus, whereas constraints and statistics both play a role in learning theories, we know little about their interaction, especially when learners are faced with multiple sources of structure.
The possibilities range along a continuum. Learners at one end could be driven primarily by statistical structure, such that they simultaneously attend to multiple types of information and weight the importance of a particular type in terms of its statistical regularity. As such, learning should closely mirror statistical structure. Thus learners encountering two types of information should favor the more statistically probable one a greater proportion of the time. I will refer to this as the statistics-driven approach. A second possibility factors a hierarchy of constraints into the picture such that learners will attend more heavily to a favored structure even if the less favored one has greater statistical certainty. This solution assumes a more minimal role for statistical processing. I will refer to this possibility as the constraints-driven approach. A possibility somewhere in the middle assumes some ordering of constraints, but further that learning will only adhere to a preferential structure to the extent that it occurs with some minimum degree of statistical certainty. Below that point learners will begin to track alternative sources of structure. Although certain kinds of information may have a more privileged status than others (because of perceptual salience, ease of processing, or prior learning), the ability to track statistical structure plays a key role in determining whether learners will focus on one type of structure or another. I will refer to this as the constraints + statistics approach.

In the next sections I will report a series of studies aimed at investigating the question of how learners choose among multiple types of structure. The first set of studies asks this question in the context of learning non-adjacent (or remote) dependencies in sequential structure.

4.4.1 Non-adjacent dependency learning

Remote dependencies in sequential structure feature in numerous high-level cognitive tasks including language (relationships between auxiliaries and inflectional morphemes as in ‘is running’ or ‘has eaten’), event knowledge (involving scripts that may have varying subevents), means–ends analysis in problem solving, and high-level planning. Such dependencies pose a considerable challenge in requiring learners to form relationships over irrelevant intervening material and are extremely difficult to acquire compared to adjacent dependencies (Newport and Aslin 2000, 2004). In contrast, adjacent dependencies appear to be privileged in terms of ease of learning or salience. As such, what might prompt learners to direct attention away from adjacent dependencies to some other form of structure?

---

2 It is well understood in the machine learning literature that learning of any great complexity is impossible unless some constraints are part of the architecture (Geman et al. 1992). Overly constrained learning mechanisms will fail to register important variations in the environment and will have difficulty generalizing. In contrast, learning mechanisms driven only by statistics will show idiosyncratic variance as a function of the training set. Thus, a key objective in understanding learning is determining what constraints are necessary and the type and degree of interaction necessary between constraints and statistics.
I first investigated this question by exposing adults to one of two artificial languages producing three-element strings (Gómez 2002). Strings were of the form aXb, cXd, and eXf in Language A (e.g. pel-kicey-jic) and aXd, bXf, and cXb in Language B (e.g. pel-kicey-rud) (Fig. 4.1). The elements a–f were pel, vot, dak, rud, jic, and tood. The 24 X-elements were wadim, kicey, puser, fengle, coomo, loga, gople, taspu, hiftam, deecha, vamey, skiger, benez, gensim, feenam, laeljeen, chila, roosa, plizet, balip, malsig, suleb, nilbo, and wiffle. Both languages contained the same adjacent dependencies, so these were not informative. Learners could only distinguish the languages by acquiring the non-adjacent dependencies. I systematically increased the size of the pool from which the middle element was drawn (set size $= 2, 6, 12, 24$) while holding the total number of strings heard during training constant to determine whether increasing variability (in the form of decreasing predictability between adjacent elements) would lead to better detection of non-adjacent dependencies (the set of 12 X-elements consisted of the first 12 words in the list, the set of six consisted of the first six, and so on). After an 18-min training phase, adult learners were tested on strings from the training language versus strings from the other language.

If learning were primarily driven by statistical structure, then learners should show a gradual increase in sensitivity to non-adjacent dependencies with an increase in set size. This is because conditional probabilities for adjacent dependencies go down as the variability of the middle element increases. However, if perceptual constraints are playing a role, learners might continue to track the privileged adjacent structure well beyond what is reasonable in terms of statistical regularity, then show an abrupt increase in sensitivity to non-adjacent structure when conditional probabilities are extremely low.

Fig. 4.1 (a) Languages used in Experiment 1 with adult participants; (b) languages used in Experiment 2 with 18-month-old infants. From Gómez R (2002). Variability and detection of invariant structure. Psychological Science, 13, 431–436. Reproduced with permission from Blackwell Publishing.
Consistent with the latter hypothesis, learners acquired non-adjacent dependencies only when the middle element was most variable, when adjacent dependencies were least predictable in set-size = 24. Accuracy across conditions in the adult study was 60 per cent (SEM = 8), 66 per cent (8.5), 65 per cent (8.5), and 90 per cent (5.5) for set sizes 2, 6, 12, and 24, respectively. In essence, performance was only slightly above chance for all but the condition with the highest variability. There was a similar pattern of results with 18-month-old infants as indicated by different listening times to strings from the training language versus the other language in a preferential listening procedure (Gómez 2002). Differences in the infant studies were that familiarization lasted for 3 min instead of 18 and infants were exposed to two non-adjacent dependencies instead of three. The infants discriminated the non-adjacent dependency when set size was 24 (mean looking time difference to grammatical versus ungrammatical strings = –2.07 s, SEM = 0.40), but not when it was 3 (mean difference = –0.04 s, SEM = 0.57) or 12 (mean difference = 0.35 s, SEM = 0.83).

One explanation for this finding is that learners focused on conditional probabilities between adjacent elements when these were relatively high (in the small set-size conditions), but when conditional probabilities between adjacent elements were sufficiently low (when set size was 24) the adjacent dependencies were no longer stable sources of structure. In this case learners focused instead on non-adjacent dependencies. Notably, learners did not show incremental increases in sensitivity to non-adjacent structure with incremental increases in variability of the middle element, a pattern we would expect if learners were responding primarily to the statistical structure. Rather they seemed to focus on adjacent dependencies long after these ceased providing reliable information. In set size 2 in the adult study, for example, the probability of the initial word being followed by a particular X-element was 0.33 (or 1 in 3). This probability decreased to 0.17, 0.08 and 0.04 with set sizes of 6, 12, and 24, respectively. In contrast, in all conditions the probability of the initial element being followed by a third element was 1. If responses to changes in conditional probabilities had been veridical, we might have expected to see learning of non-adjacent structure in the set size 6 condition (given the low conditional probability of 0.17). However, it was only after substantial variability was introduced in the middle that learners appeared to rely on the non-adjacent structure.

Subsequent research has shown that there is nothing special about a set size of 24. Gómez and Maye (in press) found that 17-month-olds can track non-adjacent structure with a set size of 18, but as in the original Gómez study, not with a set size of 12. Thus, what appears to be critical for getting learners to notice the non-adjacent dependencies is sufficient variability in the middle element where what constitutes 'sufficient' presumably varies as a function of the difficulty of the learning problem.

These results are important for increasing our understanding of how learners negotiate multiple kinds of structure and shed light on how constraints and statistical
structure both contribute to this process. They provide insight into how structure in language input can steer learning in a dynamically guided process, one that arises in the interaction of preferences that guide learning initially but that change in response to environmental pressure. The structure motivating learning appears to change as learners encounter other more stable forms of structure.

One question we might ask is whether sensitivity to different kinds of structure implies different learning mechanisms. Work with simple recurrent networks suggests that the same mechanism may be engaged in both adjacent and non-adjacent forms of learning (Elman, personal communication; Onnis et al. 2004). However, other learning problems may require learners to choose among very different kinds of structure. Thus, as a next step we might ask whether learners will switch the focus of their learning for fundamentally different kinds of structure.

4.4.2 Learning to attend to position versus co-occurrence

My students and I (Gómez et al. 2004) selected a scenario that would potentially lead to learning of one kind of structure (co-occurrence) under one set of variability conditions and of an altogether different kind of structure (position) under another set of conditions, all the while keeping information about position the same. During training, adults were exposed to auditory strings of the form aX and Yb and on a later grammaticality test they had to distinguish these strings from strings of the form bX and Ya (Fig. 4.2). Additionally, X and Y elements were drawn from a set of 1 or 24 words. There was one a-element and one b, such that in the small set-size condition participants could either focus on specific bigrams (co-occurring elements) or on the position of particular words (or both). However, when set size was large, X and Y elements varied in relationship to the more stable as and bs, and so participants might pay more attention to a- and b-words as anchor points in strings (see Valian and Coulson 1988). Importantly, participants could extract information about both co-occurrence and position in the small set-size condition because knowledge of the aX and Yb bigrams automatically yields knowledge of position. However, we were interested in determining whether learners would be more likely to
notice one form of structure over another as a function of the variability of X- and Y-elements.

Introductory Psychology students at the University of Arizona were exposed to strings from one of two artificial languages during training. Language A produced sentences of the form aX and Yb. Language B produced sentences of the form bX and Ya. The a-word was pel and the b-word was vot. Learners in set size 1 were exposed to one X- and one Y-word, and thus heard two unique strings. Those in set size 24 were exposed to the 24 Xs and 24 Ys and thus heard 48 unique strings. The frequency of a- and b-words was held constant so that, for example, learners of Language A with set size 1 heard 24 instances of pel-coomo and 24 instances of wadim-vot and learners of set size 24 heard pel preceding each of 24 X-words and vot following 24 Y-words. The strings were presented in random order in a series of blocks with training lasting approximately 9 min.

We tested learners on strings with old versus new Xs and Ys. Participants should show different patterns of responding if the differences in variability of X- and Y-elements have led them to focus on different kinds of structure. Those who have focused on co-occurrence should accurately discriminate old grammatical strings from ungrammatical ones, but should perform poorly on strings with novel X- and Y-elements. Alternately, participants who have focused on the position of a- and b-elements should perform equally well regardless of whether test strings have old or new Xs and Ys.

After acquisition, participants were told they had been listening to an artificial language that followed a set of grammatical rules. Learning was tested in two ways, by a grammaticality test and by verbal report. The type of test was counterbalanced so that half of the participants received a grammaticality test followed by verbal report (the other half were tested in the opposite order). During the grammaticality test a string was played and learners were asked to judge whether it belonged to their training language by responding ‘Yes’ or ‘No.’ We obtained verbal reports by querying learners in increasing specificity, first asking whether they noticed any patterns or rules in the strings, next whether they could report specific strings, and finally we asked them to clarify and elaborate on their answers.

We measured grammaticality performance as the percentage of correct responses to items on the grammaticality test. The data resulted in a significant two-way interaction of set size and old vs. new XYS such that learners in the set size 1 condition learned more about co-occurrence (M = 78 per cent accuracy, SEM = 1.5) than they did about position (M = 62 per cent, SEM = 2.2). Although there was some sensitivity to position (reflected in a comparison to chance [0.50]), learning was significantly greater for co-occurrence. In contrast, learners in the set size 24 conditions performed equally well when probed with strings containing old (M = 68 per cent, SEM = 1.7) versus new (M = 73 per cent, SEM = 2.3) X- and Y-elements. Furthermore, learning was significantly greater than chance.
The verbal reports were also revealing. We coded participants’ responses in terms of their knowledge of rules and patterns characterizing strings. Co-occurrence knowledge was indicated by participants’ ability to generate specific two-word phrases or by a general rule stating that there were four unique words combined as to form two two-word phrases. Participants received a verbal co-occurrence score of 1 for producing two grammatical strings or for stating the general rule, a score of 0.5 for producing one grammatical string, and a score of 0 for no verbal knowledge. Verbal knowledge of position was evidenced by stating that a particular word began strings, whereas, another ended them. Participants who stated this general rule and further specified the words used in a and b positions were given a score of 1. Those exhibiting verbal knowledge of only one a or b position were given a score of 0.5 and those exhibiting no verbal position knowledge were given a zero.

Consistent with the grammaticality judgments, verbal knowledge of co-occurrence was significantly higher for set size 1 (M = 0.73, SEM = 0.08) than for set size 24 learners (M = 0.17, SEM = 0.07). Thirty of 48 participants received the maximum score of 1 in set size 1. Six of 48 participants received the maximum score of 1 in set size 24.

Verbal knowledge of position was significantly lower for set size 1 (M = 0.02, SEM = 0.004) than for set size 24 learners (M = 0.54, SEM = 0.11). One participant (of 48) received a maximum score of 1 in set size 1, whereas 23 received the maximum score in set size 24.

In summary, verbal information generated by participants in set size 1 almost entirely reflected knowledge of co-occurrence. In contrast, verbal knowledge of position predominated in the set size 24 condition. Furthermore, seven participants in this condition spontaneously reported that the a- and b-words 'popped out,' whereas none of the participants in set size 1 reported such an effect. Although information about position was available to learners in the set size 1 condition, what was most relevant to them was the co-occurrence relationship between adjacent items. In set size 24, the relevant information was position of the a and b-elements.

The purpose of this pilot study was to determine whether variability of the X- and Y-elements would cause learners to focus on one type of structure in the stimulus over another, and to do so under conditions that would require very different forms of learning. If information about co-occurrence is privileged, then with a set size of 1 learners might focus on specific bigrams even though information about position was present. With a larger set size, however, learners might focus on anchor points in strings created by the relative stability of a- and b-elements in relationship to the changing Xs and Ys. Consistent with our hypothesis, learning appeared to be guided primarily by the most stable structure.

It could be protested that the extreme set sizes of 1 and 24 trivialized learning. After all, participants in set size 1 heard the same two strings repeatedly for a period of minutes and so it is not surprising to find learning of this information. However, the
choice of extremes was deliberate given the desire to maximize learning, especially in
the set size 1 condition where learners were more likely to acquire two kinds of
structure. Nevertheless, only one participant (of 48) in set size 1 abstracted a rule
involving anchor position, even though position should be easy to infer from memory
of specific bigrams. It appears that what people learn is entirely different in the two
conditions.

With that said, additional empirical work is in order. In particular, it will be
important to test less extreme set size manipulations. We will also want to test
increments of increasing variability between the extremes to explore how much
variability in X- and Y-elements is necessary before learners become aware of position.
How closely will sensitivity to position match the statistical decrease in conditional
probabilities that accompanies an increase in set size? Will this reflect a different
relationship between constraints and statistics than was suggested by Gómez (2002)?
It will also be important to investigate these questions with younger learners. Such
studies are currently underway.

The current study was useful for determining the conditions that might lead to
attention to position versus co-occurrence; however, both forms of structure were
concrete. Given that a critical milestone in learning is generalization, it is important to
ask what might lead learners to focus on abstract versus concrete structure. We
investigated this with 12-month-old infants in the context of form-based category
abstraction.

4.4.3 The role of variability in category-based abstraction

Abstraction is the very root of complex learning. Indeed, the generative power of
human language stems from our human ability to generalize from one instance to
another. Once a novel word is categorized, language learners can automatically apply
syntactic constraints associated with other words in its category. Given the centrality
of categorization in language, it becomes necessary to ask how children achieve this
kind of generalization.

A widely held view emphasizes the discovery of categories by first noting semantic
or referential information (Grimshaw 1981; Pinker 1984); however, it is crucial to ask
how children might begin to identify categories in the sound structure of language.
Although semantic information most certainly plays a role, prelinguistic infants are
limited in their knowledge of semantics. In contrast, they are acutely attuned to
the sound properties of language (Juszczyk 1997). If infants can identify categories in
the speech stream by means of phonological cues, they might then use this informa-
tion to learn predictive relationships between categories. In English, for example,
children must learn that function words such as ‘the’ and ‘a’ precede nouns and not
verbs, whereas ‘will’ and ‘can’ precede verbs but not nouns. Infants who have
identified categories in speech and the relationships between them will be at an
advantage with respect to the later task of mapping between meaning and form,
compared to children who only begin this process once semantic knowledge is more fully in place (Gómez and Gerken 2000).

In previous work (Gómez and Lakusta 2004), we have explored the foundations of this process by asking whether 12-month-olds would learn the relationship between functor-like a- and b-words and X- and Y-categories. During training infants were exposed to one of two training languages. One language consisted of aX and bY pairings, the other consisted of aY and bX pairs. Xs were instantiated as disyllabic words and Ys were monosyllabic. Syllable number was used as an abstract feature because information like this abounds in language. In English for instance, nouns tend to have more syllables than verbs and also tend to receive first syllable stress (Kelly 1992). Additionally, children are sensitive to these phonological cues (Cassidy and Kelly 1991). We tested infants on new phrases from their training language versus phrases from the other language. All X- and Y-words were novel at test. Infants were able to generalize to the novel sentences, suggesting that they had become sensitive to the relationships between the particular a- and b-elements and an abstract feature (syllable number). It is important to point out that although prosody is a factor (infants have to notice that X- and Y-words have either one or two syllables), this is not merely a prosodic effect. Infants had to learn to associate certain a- and b-words with particular syllable features.

I am currently using this paradigm to determine the conditions that might lead learners to focus on abstract versus concrete information by asking whether the same age infants will generalize with a smaller set size. If so, this would suggest that they are equally able to extract abstract features under conditions of low and high variability, implying that variability manipulations do not contribute to abstraction. However, if infants fail to discriminate, this suggests that abstraction is aided by exposure to a larger number of instances.

Twenty-four 12-month-old infants exposed to six Xs and six Ys (these data were originally reported in Gómez and Lakusta 2004, Experiment 1) were compared with another group exposed to a set size of 3 (Fig. 4.3). During training infants were exposed to strings from one of two training languages, A or B. Each language contained two a-elements (alt, ush) and two bs (ong and erd). In the Set-size 6 condition, there were 6 Xs (coomo, fengle, kicey, loga, paylig, wazil), and six Ys (deech, ghope, jic, skige, vabe, tam). A subset of these was used in the Set-size 3 condition so that there were three Xs and three Ys. X-elements were disyllabic and Ys were monosyllabic, thus infants could either attend to specific X- and Y-words or to the abstract feature of syllable number. The elements were combined to form grammatical phrases (e.g. alt coomo and erd deech in Language A). Strings were uttered in the same infant-directed speech with the same rising intonation structure. For each language in set-size 6 there were 24 phrases presented in random order. For Set-size 3, there were 12 phases. These were presented twice as often as in Set-size 6 to preserve the frequency of a- and b-words with the one- and two-syllable forms. Thus in both
set-size conditions the relationship between the a- and b-words and the abstract features were the same. The only difference was in the variability of the X- and Y-category members. The question was whether higher variability would aid abstraction of the category cue.

Each infant was tested using an auditory preference procedure. X- and Y-words were novel at test and thus differential listening times for legal versus illegal strings would indicate generalization of the abstract feature. Per the results of Gómez and Lakusta (2004), infants in the set size 6 condition listened longer to strings from their training language than to strings from the other language, $M = 8.44$ s (SEM = 0.72) versus $M = 7.06$ s (SEM = 0.6). Eighteen out of 24 infants showed this pattern, suggesting that they had acquired some sensitivity to the category-based structure of their training grammar. There were no differences in listening times in the set size 3 condition however, $M = 8.25$ s (SEM = 2.07) to strings from the training grammar and 9.66 s (SEM = 2.01) to the other grammar. Twelve out of 24 infants listened longer to strings from their training grammar. This is interesting in light of the fact that infants this age are extremely sensitive to prosodic structure induced by one-versus two-element words. Apparently prosody was not a strong enough cue to produce generalization to strings containing novel X- and Y-elements even though these strings maintained the prosody from training.

The ability to discriminate legal from illegal marker-feature pairings in the set size 6 condition, despite the fact that X and Y-elements were novel at test, reflects sensitivity to the co-occurrence relations between markers and X and Y categories based on their abstract features. Such learning is complex – infants had to track four markers, associate them with abstract features, and generalize to pairings containing novel words. The fact that infants were able to generalize to novel X- and Y-elements suggests that learning was abstract (involving grouping of the X- and Y-elements according to syllable number). However abstraction appears to be dependent on the amount of variability in the X- and Y-elements. Infants in the set size 3 condition showed no generalization. One interpretation is that learners in the high variability condition learned a relationship between a- and b-words and abstract features whereas learners in the low variability condition were tracking specific co-occurrence relationships. I am

![Fig. 4.3](languages.png)

Languages used to contrast learning of co-occurrence versus abstract marker-feature relationships with 12-month-old infants.
currently collecting data to determine whether learners in the set size 3 condition will distinguish strings heard during training from ungrammatical ones. This is important for establishing whether infants in this condition were tracking specific aX and bY bigrams. If so, this would add to the argument that under conditions of low variability infants focus on the co-occurrence relation between specific aX or bY pairs, whereas with enough variability they abstract higher-order prosodic features. It would also add to the argument that learning is driven by an interaction between innate sensitivities and environmental structure. Under certain conditions learners show a tendency to track co-occurrence (possibly a default sensitivity), but with enough variability in their input, they will seek out other forms of reliably cued structure. Thus, it is in the interaction of particular sensitivities (or constraints) and environmental structure that learning unfolds dynamically.

4.5 Summary

The aim of this chapter has been to summarize the literature on statistical learning, especially with regard to addressing the problem of how learners might choose among multiple kinds of structure, whether the process might be guided primarily by constraints, primarily by statistics, or whether it might arise in the interaction of constraints and statistics. In all three of the cases highlighted (adjacent vs. non-adjacent dependencies, co-occurrence vs. position, and co-occurrence vs. abstraction) we see similar patterns. Learners easily track information about co-occurrence suggesting that this may be a preferred form of structure, but with enough variability in their input, they will focus on alternate structure, suggesting how learning might be guided in the interaction of constraints and the pressures exerted by the structure learners hear.

At first glance, it might seem paradoxical that variability can aid learning. Indeed, on most accounts high variability should result in increased noise and thus decreased learning. However, high variability acts to increase the salience of alternate information, and in this way may facilitate learning. This is consistent with the idea that learning involves a tendency to seek out invariant structure, or structure remaining constant across varying contexts (EJ Gibson 1969; JJ Gibson 1966). More specifically, even though multiple types of structure were present, learners in these studies appeared to rely on one type of structure at a time. For example in the case of learning long-distance dependencies, even though the information about these was identical in all of the set-size conditions (the conditional probability of the non-adjacent dependencies was 1.0), adjacency appeared invariant when adjacent conditional probabilities were relatively high (in the small set-size conditions). It was only when adjacent conditional probabilities were sufficiently low (as when the middle element was drawn from a set of 24) that the non-adjacent dependencies stood out as invariant structure. Additionally, in the study comparing learning of co-occurrence and position, information about position of a-and b-elements was the same in the low
(set size = 1) and high variability (set size = 24) conditions. Yet, learners showed little sensitivity to position when variability was low. Finally, variability appears to play a role in abstraction. Although 12-month-old infants generalized an abstract feature (syllable number) in a high variability condition (set size = 6) they did not do so with set size = 3. Although there is still much to do in this line of research, the initial findings are intriguing. They suggest that learning arises in the interaction of perceptual sensitivities and statistical structure and suggest how structure learners encounter might guide learning.

What of Pinker’s contention that children need either to be constrained from erroneous hypothesizing or have a means for recovery when they do overgeneralize? If these studies are to be believed, there is little evidence of rampant overgeneralization in statistical learning. Instead learners appear to be conservative in their hypothesizing. In particular, in the case of non-adjacent dependencies learners seem to need a great deal of variability to reveal the invariance of the non-adjacent structure. Learners are also quite conservative about certain types of generalizations. The studies of Braine (1987), Frigo and McDonald (1998), and Gerken et al. (1999, in press) demonstrate this fact. In particular, learners will not abstract categories in an aX/bY paradigm unless they have good reason to do so, namely when cues are present for distinguishing at least some of the category structure.

In the event that children do make erroneous generalizations, how might they recover? Recent work in memory reconsolidation is relevant to this issue. According to this literature, memory is much more dynamic than was previously thought (Nader 2003). Far from being permanent, when accessed, memories are put into a labile state. Once in this state memories can be reinforced, but just as easily changed or overridden. Especially relevant to the problem of how learners might recover from overgeneralizations, Walker et al. (2003) found that knowledge of a reactivated sequence (in a sequence learning task) could be overwritten by a new sequence if exposure to the new sequence was immediate. Learners showed robust memory for a sequence after both 1- and 2-day delays, but if another sequence was presented immediately after activating the first, then memory of the first sequence was hindered. In contrast, memory of the second sequence was intact.

A similar process could be instrumental in language acquisition. If we assume that hypotheses are activated by memory, then we must also assume that engaging them puts them into a fragile state. Once activated, erroneous hypotheses become subject to change. In particular, because of the data-driven nature of learning, overly general hypotheses are likely to be replaced by more specific ones. In short, given the dynamic nature of memory and its reconsolidation processes, as well as the data driven nature of learning, it seems unlikely that learners will hold fast to erroneous or overly general hypotheses without good evidence for doing so. We are currently conducting studies investigating this hypothesis.
With this said, there are many challenges for learning approaches. One has to do with better characterizing learning processes. A particular challenge goes out to computational modelers. Although a Simple Recurrent Network has been successful at modeling acquisition of adjacent and non-adjacent dependencies (Onnis et al. 2004), the model did not exhibit the abrupt increase in performance found in the Gómez (2002) data. Instead, it showed an incremental increase in sensitivity to non-adjacent structure. Although connectionist models are capable of exhibiting non-linear behavior, this particular model did not do so. It is also important to know more about the accuracy with which learners track statistical structure. Do learners keep track of fine gradations or do they track structure by means of broad categories or thresholds? Support for the latter would be consistent with a constraints + statistics approach. There is also a great deal more to learn about the initial state. We may be able to characterize constraint hierarchies of learning (how different forms of learning are ordered in terms of salience or learnability) in terms of how easily certain types of structure are learned compared to others. Memory studies may also be informative in this respect in terms of the extent to which different forms of learning can be reinforced or disrupted. Tenacity, in terms of the amount of variability necessary for switching learners to another state, may also reveal the ordering of sensitivities in a constraint hierarchy.

In all, the present literature on learning paints a very different picture from that portrayed in traditional characterizations (Chomsky 1965; Pinker 1995). Indeed, the conservative nature of learning, combined with a dynamic memory process, conspires against traditional assumptions. Statistical learning appears to be a process akin to perceptual tuning, partially guided by constraints, but also to a great extent by environmental structure, yielding in the interaction of these pressures a dynamic, adaptive, and flexible process.

Acknowledgements

Writing of this chapter was supported by NSF BCS-0238584 and NIH R01 HD42170–01. I would like to thank Y. Munakata, R. Schvaneveldt, and two anonymous reviewers for helpful comments and suggestions.

References


