RUNNING HEAD: MODELING DISTRIBUTIONAL STATISTICAL LEARNING

iMinerva: A mathematical model of distributional statistical learning

Erik D. Thiessen

Carnegie Mellon University

Philip I. Pavlik Jr.

University of Memphis

Address for correspondence:

Erik D. Thiessen

Department of Psychology

Carnegie Mellon University

Pittsburgh, PA 15213

thiessen@andrew.cmu.edu

Abstract

Statistical learning refers to the ability to identify structure in the input based on its statistical properties. For many linguistic structures, the relevant statistical features are distributional: they are related to the frequency and variability of exemplars in the input. These distributional regularities have been suggested to play a role in many different aspects of language learning, including phonetic categories, using phonemic distinctions in word learning, and discovering non-adjacent relations. On the surface, these different aspects share few commonalities. Despite this, we demonstrate that the same computational framework can account for learning in all of these tasks. These results support two conclusions. The first is that much, and perhaps all, of distributional statistical learning can be explained by the same underlying set of processes. The second is that some aspects of language can be learned due to domain general characteristics of memory.

The term "statistical learning" is often taken to mean sensitivity to probabilistic conditional relations among sequential elements, especially in the context of discovering units in the input as in the case of word segmentation (e.g. Johnson & Seidl, 2008; Saffran, Aslin, & Newport, 1996). The ability to learn from conditional relations has been widely demonstrated across different species (e.g. Hauser, Newport, & Aslin, 2001; Toro & Trobalon, 2005) and different stimuli (e.g. Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002). Infants' sensitivity to conditional relations has attracted special attention, as this early-developing sensitivity has been suggested to play an important role in language development (e.g. Estes, Evans, Alibali, & Saffran, 2007; Hudson Kam & Newport, 2009; Misyak, Christiansen, & Tomblin, 2010; Thiessen & Saffran, 2003, 2007). However, several subsequent experiments have highlighted areas in which sensitivity to conditional relations is inadequate to acquire linguistic regularities (e.g. Endress & Bonatti, 2007; Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Toro, Nespor, Mehler, & Bonatti, 2008).

The assertion that sensitivity to conditional relations is insufficient for language learning, however, should not be taken to mean that the statistical learning approach to language development is necessarily inadequate. This is because there is a wide range of statistical relations, beyond conditional relations, which may play an important role in language learning (Thiessen, 2009). That is to say, the term "statistical learning" refers to learning from many different types of statistical information, not just conditional relations (e.g. Hunt & Aslin, 2010; Romberg & Saffran, 2010). A different class of statistical regularities can be termed distributional regularities, because they involve learning from the distributional characteristics of exemplars in the input such as

frequency and variability (e.g. Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Maye, Werker, & Gerken, 2002). It may be the case that the integration of distributional sensitivity with sensitivity to conditional relations enables statistical learning to "scale up" to the complexity of natural language (e.g. Thiessen, Kronstein, & Hufnagle, under review; Thiessen & Saffran, 2003; Thiessen & Saffran, 2007; Werker & Curtin, 2005).

Distributional statistical learning occurs in situations where learners integrate information across a set of exemplars. As such, distributional learning requires a comparison across exemplars. Comparing exemplars, and integrating information across them, yields sensitivity to the central tendency of the set. The frequency, similarity, and variability of the exemplars determine how much each exemplar contributes to the integration. An exemplar that occurs frequently will be weighted more heavily than an exemplar that occurs rarely, for example. When two exemplars are similar - a definition that depends at least in part upon the variability of the exemplars in the input set – they will tend to be integrated into the same category representation. Indeed, the contribution of distributional statistical learning to language development has largely been investigated in the context of learning to distinguish between categories, because the features that define a distribution of exemplars (frequency, similarity, variability) are highly relevant for category formation (e.g. Maye et al., 2002; Thiessen & Yee, 2010).

Despite the importance of distributional learning for language development, the process underlying sensitivity to distributional information is not completely understood. One reason for this is that the vast majority of modeling that has been done in the domain of statistical learning relates to the learning of conditional relations, especially in the context of word segmentation (e.g. Frank, Goldwater, Griffiths, & Tenenbaum, 2010; but

see Orbán, Fiser, Aslin, & Lengyel, 2008, for an example of a model of conditional learning in the visual domain; Perruchet & Vinter, 1998). While there have been fewer models exploring the contribution of distributional statistical learning to language development (though see Feldman, Griffiths, & Morgan, 2009), a comparison between the general structure of conditional and distributional statistical learning models will help to illustrate the nature of distributional statistical learning more clearly. Most recent models of conditional statistical learning seek to extract units (such as words) from the input, based on evidence that conditional statistical learning results in word-like knowledge (e.g. Estes et al., 2007; Giroux & Rey, 2009; Orbán et al., 2008). These models bind elements of the input together into a discrete representation, as when grouping syllables together into a word (e.g. Frank et al., 2010; Perruchet & Vinter, 1998).

By contrast, models that learn distributional regularities must compare *across* units (rather than binding elements together to create units) to identify the central tendency of the input (e.g. Hintzman, 1984). In models of distributional learning, previously experienced elements are synthesized or aggregated in a way that represents their central tendency (as when learners identify category boundaries or prototypes). This synthesis can lead to novel representations, as in studies of prototype formation where participants recognize objects or words they have not previously seen (e.g. Bomba & Siqueland, 1983; Endress & Mehler, 2009). Not all of these models require that an actual prototype has been formed; many store traces of individual experiences and aggregate these, rather than storing a single prototype (e.g. Hintzman, 1986; McClelland & Rumelhart, 1985). However, all of these models emphasize the central role of

integration across prior exemplars to identify central tendency. Because models of distributional learning are more concerned with integrating features to identify central tendency than they are with binding elements together into a larger representation (as in models of segmentation), models of distributional learning often assume that the input has been pre-segmented by some other process (e.g. Jusczyk, 1993). We will return to a discussion of the relation between distributional and conditional statistical learning in the General Discussion.

Prior modeling of distributional learning has largely focused on simulating prototype effects and discovery of categories in non-linguistic domains (e.g. Hintzman, 1984) (but see Clayards et al., 2008; Feldman et al., 2009). A number of statistical learning experiments demonstrate, though, that distributional learning may play an important role in identifying linguistic regularities (e.g. Gomez, 2002; Maye et al., 2002; Thiessen & Yee, 2010). Language is often held to be a unique cognitive domain that may involve specialized processes (e.g. Lidz, Gleitman, & Gleitman, 2003; Marcus, Fernandes, & Johnson, 2007). Our goal is to understand whether the domain-general principles espoused by previous models of distributional learning can also account for a wide variety of linguistically-relevant distributional learning tasks, using a single computational framework. With this goal in mind, we have developed a novel model, called "Integrative Minerva" (iMinerva for short) intended to simulate the kinds of sensitivity to the different distribution of exemplars seen in statistical learning tasks.

The iMinerva model incorporates a set of processes drawn from the theories of long-term memory: activation (of similar memories), decay, integration, and abstraction. These processes allow iMinerva to simulate a wide range of distributional learning

phenomena. In particular, we have used iMinerva to simulate three tasks where different degrees of variability in the distribution of exemplars help infants discover linguistic regularities: category learning, acquired distinctiveness, and discovery of non-adjacent relations. These simulations serve two purposes. First, they demonstrate that a relatively straightforward computational framework can account for a wide range of distributional learning. Second, they explain distributional statistical learning through the use of mechanisms of long-term memory. This type of explanation extends recent attempts to make connections between statistical learning and mechanisms of memory (e.g., Perruchet & Vinter, 1998; Thiessen, Kronstein, & Hufnagle, under review). We suggest – iMinerva is an attempt to codify this suggestion – that distributional statistical learning occurs as a consequence of the processes of long-term memory. That is to say, there is no unique "distributional learning mechanism." Rather, distributional statistical learning is a result of the way memories are encoded, recalled, and integrated with new experiences to identify commonalities.

In principle, simulating any set of distributional learning phenomena would serve to assess the claim that a single domain-general approach, based on principles of memory, is sufficient to explain distributional statistical learning. However, there are several reasons to select linguistically relevant tasks for the first assessment of this modeling approach. One is that, as mentioned previously, the majority of research on statistical learning has involved linguistic stimuli. A second is that language is a domain where domain-specific mechanisms have often been argued to be at work (e.g., Marcus et al., 2007; Lidz, Gleitman, & Gleitman, 2003), so modeling can play a particularly role in delineating the potential contributions of domain general processes in this domain.

Finally, the complexity of language ensures that there are a wide variety of linguistically relevant tasks that share very few surface commonalities. A demonstration that iMinerva can successfully simulate a diverse range of tasks provides more compelling evidence in favor of the argument that a single theoretical framework is capable of accounting for a wide range of distributional learning phenomena.

The first task we simulate with iMinerva is category learning. Maye et al. (2002) found that when infants were exposed to a unimodal distribution of phonemes along a continuum between /d/ and /t/ (where exemplars in the midpoint of the continuum occur most frequently), they failed to respond differentially to the endpoints. That is, they responded as though all of the phonemes along the continuum belonged to a single category. In this unimodal condition, infants were presented with a single distribution with a high degree of variability (that is, a relatively large standard deviation). By contrast, when infants were exposed to a bimodal distribution (where phonemes close to the endpoint are most frequent, and phonemes near the midpoint of the continuum are rare), infants responded differentially to the endpoints of the continuum. In the bimodal distribution, infants were presented with two distributions, each with a much smaller degree of variability than the unimodal distribution. The research by Maye et al. (2002) provides a clear example of infants' sensitivity to the distribution of exemplars in the input (for a conceptual replication and extension, see Maye, Weiss, & Aslin, 2008). These results suggest that the distribution of exemplars along a similarity continuum plays an important role in speech category formation (e.g. Vallabha, McClelland, Pons, Werker, & Amano, 2007).

The second phenomenon we simulate with iMinerva is the effect of the distribution of exemplars across contexts in children's use of categorical distinctions. In many word-object association tasks, children between 12 and 16 months fail to take advantage of phonemic distinctions they can hear (e.g. Pater, Stager, & Werker, 2004; Shvachkin, 1973). For example, after being habituated to an object paired with the label *daw*, children accept *taw* as a label for that object (e.g. Stager & Werker, 1997; Thiessen, 2007). Children's failure to make use of phonemic distinctions can be alleviated, however. If participants are exposed to the phonemic contrast in variable contexts (e.g., /d/ and /t/ in *dawbow* and *tawgoo*), they are more likely to use the contrast (e.g. Thiessen, 2007; Thiessen & Yee, 2010). When infants are exposed to the phonemic contrast in invariant contexts (e.g., /d/ and /t/ in *dawgoo* and *tawgoo*), they show no gain in their ability to make use of the phonemic contrast. The fact that variable contexts facilitate the use of the contrast may be related to the phenomenon of acquired distinctiveness. When two similar stimuli (A and B) are paired with two different outcomes (X and Y, forming the compound stimuli AX and BY), the similar stimuli become easier to differentiate (e.g. Hall & Honey, 1989). These results indicate that learners are sensitive to the contexts in which different stimuli are distributed, and that those distributions can help to make the distinction between exemplars from different categories more robust. A more robust distinction among the phonemic categories should make it easier to subsequently map the categories onto different referents, because discrimination is a necessary precursor to learning separate mappings (Gibson, 1940).

The third effect we simulate with iMinerva is also related to the effect of variability on learning. Prior research demonstrates that variability plays a role in

infants' ability to detect non-adjacent relations. Many of the dependencies in language are non-adjacent dependencies, as between the morphemes *is* and *ing* in phrases like *is walking* and *is running*. The intervening material between *is* and *ing* is unrelated to the dependency (any regular verb can occur between them), and thus variable. Discovering non-adjacent relations is more difficult than discovering adjacent relations (e.g., Creel, Newport, & Aslin, 2004). Fortunately, the variability of intervening elements between the non-adjacent elements can help infants to discover the non-adjacent regularity. When infants are exposed to an artificial language with non-adjacent regularities, as in the string AXB (where X is a variable element), infants discover the non-adjacent regularity between A and B more easily when the X element is more variable (Gomez, 2002). This is a striking result, because increasing variability in the X element actually presents the infants with more complex input.

As these results indicate, infants are able to take advantage of the distribution of exemplars in the input in a wide variety of learning tasks that relate to language development. But the very breadth of distributional learning raises an important question about whether all of these aspects of learning can be accomplished by the same underlying mechanisms. This is a question that has been explored with respect to *conditional* statistical learning. Humans detect conditional relations for both sequential auditory stimuli (e.g., syllables in a word) and simultaneously presented visual stimuli, such as elements of a visual array (e.g. Fiser & Aslin, 2002). Learning in these different domains is potentially mediated by different mechanisms, a hypothesis that has been assessed by both behavioral and computational research (e.g. Conway & Christiansen, 2005; Kirkham et al., 2002; Orbán et al., 2008). Computational work provides an

especially good technique for testing hypotheses about underlying mechanisms, because models allow researchers to provide an existence proof that a single hypothesized learning mechanism can account for several different kinds of learning (e.g. Seidenberg & McClelland, 1989).

However, there have been relatively few models investigating distributional learning with the kinds of linguistic stimuli often used in statistical learning tasks. The few prior models in this area are intended to explore distributional learning do so only for a single kind of task (e.g. Feldman et al., 2009). Such models are informative, but do not attempt to determine whether a single set of underlying processes can account for two or more learning phenomena. If different aspects of distributional learning are accomplished via the same processes, it will be possible to create a single model that can simulate many different aspects of distributional learning. While computational modeling cannot provide definitive evidence that a particular process underlies human learning, it can provide an existence proof that it is possible for a hypothesized process to do so.

We propose that distributional statistical learning is accomplished by processes of long-term memory, including similarity-based activation of prior memories, strength-based learning of features, abstraction of irrelevant features and memory decay. This proposal suggests that the same underlying processes are responsible for infants' success in all of the distributional learning tasks discussed above, and possibly many other forms of distributional learning. To assess this hypothesis, we simulated learning in all three tasks using the iMinerva model. This model, which shares many principles in common with the MINERVA 2 (Hintzman, 1984) model, is a model of learning from exemplars

stored in long-term memory. In a series of simulations, iMinerva was able to mimic three different aspects of distributional learning: category learning (e.g. Maye et al., 2002), acquired distinctiveness (e.g. Hall & Honey, 1989), and the facilitative effect of variability in discovering non-adjacent regularities (e.g. Gomez, 2002). This provides an existence proof that the processes of long-term memory can, in principle, account for several different aspects of distributional learning over the kinds of linguistic stimuli used in statistical learning tasks. In particular, we propose that to benefit from the characteristics of the distribution of exemplars in the input, a learner must also be able to *integrate across* exemplars over the course of learning. Only by comparing across exemplars, and integrating the current exemplar with prior experience, is it possible to learn from the central tendency and variability of exemplars in the input. Thus, the current model accounts for distributional statistical learning by utilizing a process of comparison across experienced exemplars.

The iMinerva Model

Unlike most other models of statistical learning, this model is not intended to simulate processes via which learners detect sequential conditional relations in the input (as when they detect that syllables 'go together' in a word). For the current simulations, we assume that processes not invoked by iMinerva are responsible for segmenting the input. These segmentations provide the exemplars over which iMinerva operates. Rather than segmenting the input, iMinerva is intended to simulate learning from the distribution of exemplars, especially learning that is related to the variability with which different exemplars occur. To do so, iMinerva relies on four interrelated processes: similarity of previously stored exemplars, integration of the current exemplar with previous

experience, decay of old exemplars, and abstraction from the exemplars. Together, these processes allow iMinerva to produce the entire range of distributional learning discussed above. For a complete mathematical description of the model, see the Appendix. Here, we will provide a verbal description of the key characteristics of the model.

Like all exemplar memory models, iMinerva stores prior experiences in the form of discrete exemplars. These exemplars are coded as n-dimensional vectors with positive and negative feature values. Each feature is linked to some psychologically real characteristic of the stimulus. For example, a vector describing a shape might have a features linked to the presence or absence of the characteristics "round," "three-sided," "four-sided" and "five-sided." The valence of the features describes whether the characteristic is present, absent, or contradicted (i.e., a square would have a 0 feature value for the characteristic "round", because roundness is absent, a positive feature value for the characteristic "4-sided," and a negative feature value for the characteristics "3-sided" and "5-sided," because having exactly four sides contradicts having three or five sides). The magnitude of the feature describes the model's certainty that the characteristic is present, absent, or contradicted. Feature values can, in principle, range from positive infinity to negative infinity. The greater the absolute value of the feature, the more certain the model is about the presence (or negation, for features with negative values) of the characteristic described by that feature.

Four processes drawn from research on memory – similarity-based comparison, decay, integration, and abstraction – allow iMinerva to learn from exposure to exemplars. Similarity between exemplars is computed as the cosine similarity of the two vectors on a feature by feature basis, considering both the valence and magnitude of each feature.

Note that the magnitude of a feature can change even after it has been stored in memory, due to the effect of decay. All vectors stored in memory are subject to continuous decay effects causing feature values to tend toward zero. When a new vector is presented to the model, prior exemplars in long-term memory are activated as a function of their similarity to the current exemplar. Activation is the cube of raw cosine similarity, which means that only highly similar exemplars are strongly activated.

Similarity of prior experiences causes an integration between current and prior information we have termed "engagement," where the current experience is interpreted in light of prior experience. In engagement, the current example is integrated with a similar prior exemplar to create an interpretation of the current experience. This leads to the storage of three vectors in memory: the current example, the prior vector drawn from memory, and a new interpretation. If multiple prior exemplars are similar, the model selects the strongest of them above the similarity threshold (a parameter that varies between 0 and 1.0). Interpretations are created through an additive merging of the current exemplar and the strongest prior exemplar above the similarity threshold. If the current vector has features consistent with the features of the vector (above threshold) in memory, then the engagement strengthens these features, and the resulting interpretation has more extreme feature values than either of them. If the current vector has features that are inconsistent with the features of the strongest vector (i.e., has feature values in the opposite direction), then the interpretation resulting from their engagement will have less extreme values for these features than it did previously. The additive integration of the current vector and the strongest vector in memory is controlled by the learning rate of the model, a parameter that can be set to any value greater than 0 (though values between

0 and 1 are most plausible). If the learning rate is set to a high value, the current exemplar has greater influence on the new vector that arises from the engagement process.

Interpretations are stored in the same manner as exemplars in the model, so once they are formed they can be engaged by subsequent exemplars. The creation and storage of these interpretations (i.e., engagement) is the process by which iMinerva learns the central tendency of exemplars in the input. As an example, consider what would occur if iMinerva were exposed to a series of two-feature vectors, where the first feature of all of the variables was 1, and the second feature of the vectors alternated between -1 and 1. Across a series of engagements, the first feature would be reinforced, and increase in magnitude (for example, with a learning rate of .1, the engagement between the first and second vector would produce a new interpretation with a first featural value of 1.1). In contrast, the second feature will decrease, since the learning is a function of the feature value of the new vector, -1, multiplied by the learning rate, so it will become .9. Because of this decrease, and because of forgetting, the second feature will trend to 0 over time. In this way, iMinerva continually refines its interpretations in a way that is consistent with the central tendency of the input.

The final process that is necessary for iMinerva's learning is abstraction, which facilitates generalization to novel stimuli (e.g., McClelland & Plaut, 1999). To simulate abstraction, iMinerva transforms features to null values when an interpretation contains features whose values fall below some fraction (controlled by a parameter varying between 0 and 1) of the average absolute feature strength for that interpretation. Features that are nullified in this manner are no longer used to compute similarity ratings; they

neither make a vector more nor less similar to some other vector (a vector of 2, 1, null would be equally similar to a vector with features 2, 1, -1, and a vector with features 2, 1, 3). The abstraction process was added to the model to simulate the fact that experience often results in a decrease in sensitivity to certain features of the input (e.g. Werker & Tees, 1984). Nullification of features does not necessarily mean that the features are no longer detected, but rather that they have lost salience. This allows attention to be devoted to those features that prior experience indicates are informative. Such efficient use of attention is useful both because attentional resources are limited (e.g. Miller, 1956), and also because it leads to more efficient processing of subsequent experiences (e.g. Winkielman, Halberstadt, Fazendeiro, & Catty, 2006).

The four processes invoked by iMinerva – similarity-based activation, engagement, decay, and abstraction – are drawn from research on human memory. While each process is independent, their interaction is crucial to a complete attempt to simulate human memory. For example, while engagement could occur (randomly) in the absence of information about similarity, it is the interaction of these two processes that allows iMinerva to learn in a principled fashion. Similarly, while similarity could be processed in the absence of abstraction, the presence of this process ensures that spurious relations that occasionally occur in the input do not persist in memory. That is, abstraction serves to "sharpen" iMinerva's sensitivity to regularity. Decay for its part adds realism to the model since it limits the number of active traces (since traces decay away) and captures the assumption that memory does not have unlimited capacity. Taken together, these four processes present an attempt to simulate the critical processes of human memory that we believe are responsible for distributional statistical learning.

Much of the theoretical groundwork and delineation of these processes was originally set forth in the MINERVA 2 architecture (Hintzman, 1984), from which iMinerva is adapted. However, there are important differences between the ways the two models simulate memory processes. First, in MINERVA 2, vector feature values are limited to a range between -1 and 1, whereas feature magnitude is potentially limitless in iMinerva. This means that similarity in iMinerva is influenced by both the direction and magnitude (i.e., certainty) of a feature, while similarity in MINERVA 2 is driven more by valence. Both iMinerva and MINERVA 2 are sensitive to the central tendency of the vectors stored in memory. However, they achieve this sensitivity through different mechanisms. MINERVA 2 creates weighted (by similarity) average of all of the vectors in long-term memory. Because of this, two equally similar vectors contribute equally to the overall summation, regardless of when they were originally experienced. In iMinerva, the distribution in memory is reflected in the amplitude of features in the interpretation resulting from engagement between a current experience and a prior experience. The similarity threshold means that some vectors do not contribute, even weakly, to the model's representation of central tendency. We believe that this is a more psychologically plausible assumption about the processes of recall than the approach implemented in MINERVA 2. Learning and judgments within a domain or category often reflect only the characteristics of that domain or category, rather than the characteristics of all prior experiences. The existence of a similarity threshold allows iMinerva to simulate this specificity more easily than the MINERVA 2 architecture on which it is based. Finally, discovery of regularities in iMinerva should be stronger in some cases than in MINERVA 2, due to the presence of a process of abstraction that

removes patterns from memory if they are only rarely experienced (a point we will discuss in more detail in Experiment 3).

Simulation 1

Infants begin to discover the phonemic categories of their native language even before they are familiar with the meaning of words (e.g. Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Werker & Tees, 1984). One of the features of the linguistic input that may allow them to do so is the distribution of phonetic exemplars in the input. Because sounds that fall between two categories are ambiguous, it may be the case that speakers produce exemplars in these ambiguous regions less often. If so, regions of phonetic space where relatively few exemplars (in comparison to nearby regions) are produced would correspond to boundaries between phonemic categories. While the distribution of phonetic exemplars is necessarily noisy due to individual and contextual differences in articulation, research indicates that this kind of distributional information is available in the input (e.g. Kuhl et al., 1997; Vallabha et al., 2007).

Of course, this distributional information is only useful if infants can benefit from it. To assess whether they can do so, Maye et al. (2002) presented 8-month-old infants with a simplified version of the distributional information relevant to discovering category boundaries. In their experiment, infants were exposed to phonemes along an eight-step continuum from /da/ to (unaspirated) /ta/. One group of infants was exposed to a unimodal distribution, with stimuli 4 and 5 (the two central stimuli, halfway between extreme /da/ and extreme /ta/) occurring most frequently. The other group was exposed to a bimodal distribution, where stimuli 2 and 7 (near the endpoints of the distribution) occurred most frequently, and stimuli in the middle of the continuum occurred rarely.

This bimodal distribution mimics a linguistic system with two categories, while the unimodal distribution suggests a single category. After exposure, only infants exposed to the bimodal distribution responded differentially to /da/ and /ta/ during test trials.

The goal of this simulation is to reproduce the pattern of data found by Maye et al. (2002). We hypothesize that exposure to the unimodal distribution will promote formation of a single broad representation, one that includes a wide variety of exemplars. By contrast, we predict that exposure to the bimodal distribution will promote the formation of two less inclusive representations. This would provide an explanation for the pattern of results found by Maye et al. If the model (or the infant) has only a single, relatively broad category representation, exemplars from either end of the continuum can fall into the same category. But if the model (or infant) has formed two categories, exemplars from each endpoint will fall into different categories.

Method

Our goal for the model was to show that given the bimodal or unimodal distribution of phonetic input, different representations will be formed that account for the different behaviors in response to these distributions of input. For this simulation, we used the exact distribution of stimuli shown by Maye, Werker and Gerken (2002) to produce sensitivity experimentally. In this experiment, infants were presented with 64 examples sequentially from either a bimodal or modal distribution. Table 1 shows the coding of the 16 features that were used to represent each vector from phoneme da4 (the most extreme da) to ta4 (the most extreme ta). Note that these features are drawn from linguistic tradition suggesting that acoustic input can be decomposed into binary feature arrays. This does not represent a theoretical commitment to abstract phonemic

representations, but rather an attempt to capture the similarity structure of the input in a way that is mathematically convenient for iMinerva's vector input.

Insert Table 1 about here

To illustrate this point, consider the first feature of the vectors, corresponding to voicing. This coding system captures a set of assumptions about how infants encode their experience. Recall that the magnitude of a feature in iMinerva does not (directly) reflect voice onset time (VOT). Rather, magnitude reflects the learner's confidence that the feature is present, and magnitude systematically increases for more extreme VOTs based on the assumption that more extreme values of VOT (i.e., further from the ambiguous middle tokens along the continuum) are more easily perceived as exemplars of either voicing or voicelessness. This coding scheme requires that infants be able to detect within-category variation; that is, to be able to perceive the difference between two exemplars of the same phoneme. While early models of categorical perception suggested that this was not the case, more recent research with infants indicates that infants do perceive within-category variation (e.g., McMurray & Aslin, 2005). Because of the way this coding scheme is devised, iMinerva does not address the problem of how infants learn to map a continuous acoustic characteristic (such as VOT) onto a more discrete representation (such as voicing). Rather, the simulation addresses the question posed by research of Maye et al. (2002): given that infants can perceive a difference between voiced and voiceless exemplars, why does the distributional information in the input lead them to respond to these exemplars as though they are members of the same category

(i.e., to respond equivalently to perceptually distinguishable inputs for some kinds of distributions, and lead them to respond to the exemplars as though they are members of different categories for other kinds of distributions?

While infants' response to phonetic exemplars was measured by looking time in the experiment that produced this data (Maye et al., 2002), we were more interested in producing a model that explained *why* looking time was different rather than a model that produced exacting fits to the looking time data of individual infants. For this reason, the results for the simulation are characterizations of student representations (or interpretations) that are thought to drive the difference in listening times. In this, and the later simulations in this paper, we assume that infants' gaze durations are caused by novelty and conflict. Early during habituation longer listening times are caused by the novelty of the input not matching any recent input, which drives attention and learning. Later on in learning, when new exemplars are presented they may match one or more representations in memory. If multiple representations are matched, causing conflict, we propose that there is a competitive process of selecting the best match. This idea that multiple representations cause conflict and longer latencies is a well-established idea with a history of research behind it (e.g. Van Rijn & Anderson, 2003).

Therefore, what this simulation hopes to establish is that the unimodal and bimodal frequency distributions cause different states of memory. In the case of the unimodal distribution input, the model should display a single interpretation that would not cause competition and lacks novelty because of more repetition. By contrast, in the case of the bimodal distribution, we expect to see two interpretations that would cause competition and have greater novelty due to proportionally less reinforcement.

Results

Figure 1 shows the typical pattern of memory from simulating a child in each condition. In the top graph of the figure the memories formed from bimodal input are shown, and in the bottom graph memories formed from unimodal input are shown. The x-axis (trace index) describes the order that examples and interpretations were added over time, starting at trace index 0 (the first trace). The y-axis graphs the strength of the initial feature of the vector for created examples and interpretations, because the initial feature (which reflects voicing status) was the key feature differentiating /t/ and /d/ in Maye et al (2002) research. The dashed lines connect examples to descendant interpretations. The solid lines connect the interpretations to any descendant interpretations. Of course, all of these solid lines trace back to the creation of some original interpretation from the engagement of two similar examples. Figure 1 shows the critical feature development over cycles of engagement. As we can see from the top part of the figure, when input is bimodal, two interpretations are formed early and strengthened with the engagements of multiple examples. The bottom of the figure shows how only a single interpretation is formed when input is unimodal. Since it is impossible to graph the NA values for the abstraction in the single interpretation we plotted the solid line at a value of exactly 0 (*to represent null in a way that would still be visible*) in Figure 1 bottom.

Insert Figure 1 about here

Because learners in Maye et al's (2002) experiment received the stimuli in random order, and because we noted that very occasionally the bimodal input would

result in the formation of only a single interpretation, it was important to run the model multiple times. This would ensure that while showing some individual difference (which occurs in children as well) the basic pattern of results showed a strong difference for the two input distributions. Figure 2 shows this comparison for 500 simulated children. From these results it is clear that the bimodal input consistently produces 2 interpretations, while the unimodal input produces one representation. Unique interpretations were determined by tracing forward from each initial interpretation to find the most recent version of the interpretation. The few exceptional children in each condition randomly received orders of practice that allowed them to initially entrench an inaccurate interpretation or interpretations of the input stimuli distribution, which was (were) subsequently strengthened well enough to persist. The fact that our simulation shows these individual differences helps illustrate that our process-based explanation is not just a statistical summary of the results, but rather a mechanistic account of how children process the input to arrive at personal interpretations of their experiences. While it did not have much effect on the result for this experiment because of the variable order of input, we added some variability (random normal with a SD of 0.05) to the threshold for each simulated child, so as to insure that our result was not somehow due to assuming our simulated subjects had no individual differences. This variability also means that some simulated learners are biased to form more or fewer interpretations.

While we did not simulate looking times, the representations that the model develops in response to the bimodal and unimodal exposure are consistent with Maye et al's (2002) data. In that experiment, infants were exposed to two kinds of test trials: alternating (between ta and da stimuli) and non-alternating trials. Only infants in the

bimodal condition were able to discriminate between alternating and non-alternating trials. Those models exposed to unimodal input would similarly have difficulty distinguishing between alternating and non-alternating trials, because both kinds of stimuli (ta and da) yield the same interpretation. Because of this, alternating trials would provoke the same interpretation as non-alternating trials. For models exposed to the bimodal input, alternating trials would invoke two distinct interpretations, which is a different pattern than the single interpretation invoked by non-alternating trials.

Insert Figure 2 about here

The representations formed by the model in the unimodal and bimodal cases differ primarily on the feature representing voicing, as would be expected from the input. In the bimodal condition, the model's interpretation in response to voiced and voiceless test items has an absolute value of about 2 for the feature representing voicing. That is, the magnitude (i.e., certainty) of the voicing feature has increased as a function of training. After exposure to the bimodal input, iMinerva becomes even more confident in the voicing distinction that occurs in the input. In the unimodal condition, iMinerva creates a single strong representation. But rather than retaining the ta-da difference in this representation, the model actually abstracts away this feature. This was generally confirmed by comparing the total percent abstracted (null) initial feature in the unimodal condition to other features values in the final interpretations counts. For the unimodal condition we found 85.4% of the interpretations had a null initial feature, while in the bimodal condition only 6.36% of the interpretations were null for the initial feature. In

other words, the unimodal distribution of input led to simulated children discounting the salience of the /t/-/d/ distinction due to the distribution of input, while the bimodal condition strengthened the model's confidence that the distinction was occurring.

It is important to note that iMinerva is not the first model to demonstrate that distributional features of the input can help infants adapt to the phonemic structure of the input. Many computational architectures – including connectionist and Bayesian models, in addition to exemplar memory models – are capable of learning from this kind of distributional information (e.g., Feldman et al., 2009; McMurray, Aslin, & Toscano, 2009; Vallabha et al., 2007). Indeed, many of these models are superior to iMinerva in that they learn from linguistic input that is closer to the complexity of natural language, or have representational schemes that more closely reflect acoustic features of the input. Our goal here is not to argue for the superiority of the iMinerva architecture in simulating this (or any) particular aspect of distributional learning. Indeed, it may be the case that all of these models are merely different computational instantiations of the same underlying psychological processes (e.g., Shi, Griffiths, Feldman, & Sanborn, 2010). Rather, our goal is to demonstrate that iMinerva is able to account for many different distributional learning phenomena within a unified approach, thereby providing an existence proof that all of these phenomena – though distinct on the surface – can be explained by the same set of processes. To this end, we will next use iMinerva to simulate a different distributional learning problem.

Simulation 2

By the end of the first year of life, infants have made much progress toward identifying the phonemic categories of their native language (e.g. Kuhl et al., 2006;

Werker & Tees, 1984). However, in some settings they may not use those categories in an adult-like manner. For example, in habituation tasks where infants learn novel word-object associations, 14-month-olds treat minimal pair labels (such as *daw* and *taw*) as interchangeable labels for the same object (e.g. Stager & Werker, 1997). By 17-20 months, infants improve in this task, and now correctly reject minimal pairs as labels for an object they have previously seen labeled with a different label. At 17 months, the ability to correctly reject a minimal pair is linked to vocabulary size (e.g. Werker, Fennell, Corcoran, & Stager, 2002). Infants with larger vocabularies are more likely to succeed, suggesting that experience with the distribution of phonemes in lexical contexts plays a role in the ability to use those phonemes (e.g. Thiessen, 2007).

To test this hypothesis, Thiessen (2007) provided 15-month-olds with exposure to phonemes in variable lexical contexts (such as /d/ and /t/ in *dawbow* and *tawgoo*). Embedding the phonemes in distinct contexts mimics the distributional characteristics to which children are exposed as they acquire their native language. Unlike adults, children know very few words were phonemes occur in minimal pairs (as in *deer* and *tear*). Instead, children are likely to be familiar with words where phonemes occur in distinct contexts (e.g. Caselli et al., 1995). The results of the experiment indicated that familiarization with *dawbow* and *tawgoo* facilitated use of the contrast between *daw* and *taw*. After exposure to *dawbow* and *tawgoo* (which provide examples of the phonemes /d/ and /t/ in different contexts), children no longer treated *daw* and *taw* as interchangeable (Thiessen, 2007; Thiessen & Yee, 2010). This effect was not simply due to greater familiarity with the sounds *daw* and *taw*. A separate group of children were familiarized with *dawgoo* and *tawgoo* (which provide examples of the phonemes /d/ and

/t/ in identical contexts) and showed no facilitation. These results suggest that experiencing the phonemes in different contexts facilitates their use.

The goal of this simulation is to replicate the results of Thiessen (2007). We hypothesize that exposure to the phonemes /d/ and /t/ in different contexts will yield divergent representations. That is, when the model is presented with "daw," this will activate prior experiences with the phoneme /d/ in the context of "dawbow." When the model is presented with "taw," this will activate prior experiences with the phoneme /t/ in the context of "tawgoo." This will lead to divergent interpretations of the phonemes, and make the distinction between them more robust. By contrast, if the model experiences the phonemes in identical contexts (like /d/ and /t/ in "dawgoo" and "tawgoo"), the interpretations of /d/ and /t/ will be convergent.

Method

To simulate the input in Thiessen (2007), we were able to reuse the da4 and ta4 stimuli vectors from Simulation 1 (Table 1), because da4 corresponds to "daw" and ta4 corresponds to "taw". In addition to the these phonemes, which were represented with 16 features again, we also had 2 suffixes, "bow" and "goo" which were also represented with 16 features (these are shown in Table 2). This meant that for each of the Thiessen stimuli we required a 32 feature vector except for "daw" and "taw". For "daw" and "taw" we choose to pad the stimuli vector with 16 null values to represent the fact that there was nothing to compare for this portion of the stimuli unit. According to our similarity function (see Appendix), comparing a feature to a null feature causes no effect on the similarity. This corresponds to an assumption that similarity is a partial matching

process that can cope with sparse information, and in such cases, simply reports back based on the comparison of the information available.

Insert Table 2 about here

Again in the Thiessen (2007) case we were interested in showing how the two different input streams caused interpretation states that resulted in discrimination of daw and taw. Because we were interested in retaining as much commonality as possible between the 3 simulations in this paper, we only varied the threshold of similarity in this simulation with respect to Simulation 1. We lowered it from a mean 0.85 from the Maye simulation to 0.6. Lowering it is consistent with the idea we might expect a greater propensity for noticing similarity in these older children (e.g. Cohen, 1991, 1998).

Results

Figure 3 shows the interpretation counts in the Thiessen (2007) condition simulations. It is clear that the simulation is consistently forming two interpretations (99.4% of the time) for the "daw, dawbow, tawgoo" condition (henceforth: the distinct contexts condition) and one interpretation (87.6% of the time) for the "daw, dawgoo, tawgoo" (henceforth: the identical contexts) condition. Looking in more detail at the feature vectors for the interpretations revealed that in the distinct contexts case, we found that "daw" and "dawbow" formed a single interpretation with some of the "bow" phoneme moderately included, while "tawgoo" formed its own weaker interpretation. In the identical contexts condition, one interpretation is formed that has a weak /d/ - /t/

feature, maximal support for the “aw” features, and moderate support for the “goo” features.

Insert Figure 3 about here

Infants in the Thiessen (2007) experiment distinguished between “daw” and “taw” test trials after exposure to “dawbow” and “tawgoo;” they failed to do so after exposure to “dawgoo” and “tawgoo.” As in Simulation 1, this result is consistent with the interpretations formed by the model after the different exposure regimens. Figure 4 shows how these final interpretations are represented in the model. The models exposed to “dawgoo” and “tawgoo” (i.e., the identical contexts) form a single interpretation that both “daw” and “taw” activate. That is, these models fail to differentiate between “daw” and “taw” in much the same way as infants exposed to “dawgoo” and “tawgoo.” By contrast, the models exposed to “dawbow” and “tawgoo” (the distinct contexts) create different interpretations of “daw” and “taw.” Due to their experience with “daw” and “taw” in different contexts, these models have the capability to respond differentially to the two syllables.

Insert Figure 4 about here

As in Simulation 1, we inspected the representations iMinerva formed to better understand why the model performed differently across conditions. The key difference across conditions is that in the identical contexts condition, both test items give rise to an

interpretation that is identical on the second syllable. In the distinct contexts condition, the test items yield interpretations that differ on their second syllable: "daw" activates memories of "dawbow," while "taw" activates memories of "tawgoo." Importantly, iMinerva does not predict that infants actually *perceive* the presence of a second syllable after the presentation of the two test items. Rather, the interpretations indicate that infants *recall* (likely implicitly) that they have seen these two syllables presented in different contexts; that is, the two test trials activate different sets of memories. This perspective is consistent with many demonstrations that implicit memory can influence behavior, even in the absence of conscious awareness.

In both Simulation 1 and Simulation 2, iMinerva succeeds because the distributional characteristics of the input (the frequency of exemplars along a continuum, or the lexical context in which a phoneme occurs) alter the interpretation of subsequent input. This commonality reflects our claim that the same set of processes – instantiated in iMinerva – can explain a wide variety of distributional learning. As a point of comparison, let us briefly consider some of the other models that have been suggested to explain the kinds of distributional learning that we have used iMinerva to simulate. We suggest that an advantage of iMinerva is its flexibility. For example, the kinds of models that are capable of learning categories from phonetic distributions akin – though more complex – to those in Simulation 1 (e.g., McMurray et al., 2009; Vallabha et al., 2007) have not been applied to the acquired distinctiveness learning iMinerva demonstrates in Simulation 2.

One exception to this is the model developed by Feldman et al. (2009), which extends the prior models distributional phonemic learning by adopting a Bayesian

approach that is also capable of learning from the distribution of phonemes in lexical across lexical contexts. This model, like iMinerva, is capable of learning in the kinds of tasks in both Simulation 1 and Simulation 2. It differs from iMinerva in that, as a Bayesian model, it is not an implementation of psychological process, but rather an exploration of the kinds of structure that can benefit an optimal learner. Additionally, the Feldman et al. model is limited in that its learning is limited – based on the hypotheses it considers and the structure of the model – to phonemic and lexical learning. As we will demonstrate in Simulation 3, iMinerva can be extended even beyond these domains of learning to simulate a distributional learning phenomenon to which no prior model has been applied.

Simulation 3

Discovering non-adjacent relations is critically important for language development. This is due to the fact that syntactic patterns are often organized hierarchically, rather than obeying adjacent regularities (e.g. Chomsky, 1959). For example, in a noun phrase, the article *the* signals that a noun will occur, but the noun can occur several words later (as in *the big shaggy dog*). A variety of experiments have demonstrated that infants and adults are capable of detecting these kinds of non-adjacent regularities in artificial languages (e.g. Gomez, 2002; Mintz, 2002, 2003). Indeed, in statistical learning tasks, it is even possible to segment words based on non-adjacent regularities, as when the first syllable predicts the third syllable but not the intervening second syllable (Creel, Newport, & Aslin, 2004; Newport & Aslin, 2004).

While infants and adults are capable of detecting non-adjacent relations, they appear to be more difficult to learn than adjacent relations, and in some cases may only

be detected if there is some perceptual or structural cue highlighting their existence (Creel et al., 2004). One structural cue that helps to highlight non-adjacent relations is variability. When exposed to a non-adjacent regularity such as A-X-B (where the A element predicts the B element, and the intervening element is variable), the variability of the intervening X element is critical. Gomez (2002) found that infants failed to detect the non-adjacent A-B relation if there were 3 or 12 possible intervening X elements. Infants only succeeded when there were 24 possible intervening X elements. This success is striking, as increasing the number of X elements actually makes the input more complex. Our goal in this simulation is to understand why variability helps infants to detect relations among the less variable elements of the input.

Although the effect of variability on learning is a striking and robust phenomenon (e.g. Gomez & Maye, 2001), the mechanisms underlying the facilitative effect of variability are not completely understood. Indeed, some models of statistical learning are unable to detect these kinds of non-adjacent relations. Chunking models, for example, simulate statistical learning by forming discrete, unitized representations of adjacent syllables (e.g. Perruchet & Vinter, 1998). By definition, these kinds of models are unable to detect non-adjacent relations (e.g. Perruchet, Tyler, Galland, & Peereman, 2004). We propose that abstraction plays a central role in the ability to benefit from variability. That is, when infants are presented with A-X-B strings with only a few surface forms (i.e., few possible intervening X elements), the X elements are represented in memory. But when there are many possible intervening elements, only the A and B elements are represented because the individual possible X elements will contradict each other to the point that the features representing X are abstracted. Based on this, we hypothesize that learning the

benefit of high variability will be strongest when our abstraction parameter is set to a relatively high value, and that variability will be less beneficial when the abstraction parameter is set to a lower value.

Method

The goal of the Gomez simulation was to simulate how children may have learned the two non-adjacent patterns by forming two strong interpretations, one corresponding to each pattern. As in simulations 1 and 2, we used the same 16 feature vector per phoneme representation of the stimuli (see Appendix). Gomez used two different non-adjacent regularities for each of the three conditions, A-X-B and C-X-D, in which the intervening X was either 3, 12 or 24 different items. In order to equalize quantity of practice, this meant that there were either 6 stimuli (3 intervening x 2 intervened between) repeated 8 times in random order, 24 stimuli (12 intervening x 2 intervened between) repeated twice in in random order, or 48 stimuli (24 intervening x 2 intervened between) repeated once in random order.

Results

Figure 5 shows the result after 500 simulated children in each condition. Displayed is the percentage of learners that have the specified number of representations above the criterions of 0.75 (left side) and 2.0 (right side). Numbers on the left are greater than numbers on the right because, with the higher criterion, fewer interpretations per student are counted. We show these results for two different criteria to highlight that while the model produces similar patterns for the count of strong interpretations (above 2.0) per simulated child, for 3 and 12 intervening items the model is making qualitatively different predictions about the interpretations learned (as revealed by the lower criterion

figures). For example, for 3 intervening pairs, we see that either 4 or 6 interpretations are formed per simulated child by looking at the percentages above the lower (0.75) threshold. These 6 interpretations are easily explained since this condition repeated 6 stimuli, so it is hardly surprising that 6 verbatim interpretations were formed. Further, we can see that occasionally 4 interpretations are formed. Inspecting the interpretation vectors shows that this happens if a particular simulated student's presentation order or lower threshold allows it to group 3 of the 6 stimuli (those belonging to either the 1st or the 2nd intervened between phoneme pair), leaving the other 3 ungrouped. When this grouping begins, 5 interpretations become unlikely because the grouping of 2 exemplars with the same intervened between phonemes draws in the similar third intervened between pair in the set.

Insert Figure 5 about here

In contrast, in the 12 intervening pairs case, the intervening X phonemes are repeated four times less often when compared to the three intervening pairs case, but there is still some repetition on the intervening phonemes, and this causes incorrect abstractions and interpretations that block proper noticing of the dependency between first and last syllables. It is noteworthy, that unlike the three intervening item case, occasionally this model does form two strong interpretations, but by comparing panels b, d, and f in Figure 5 we cans see that 12 intervening items is only very slightly better than three intervening item case, while 24 intervening items is about twice as good at forming 2 strong interpretations (>2 mean feature strength). This categorical difference between

the conditions accounts for the pattern of results in Gomez (2002). When the model only forms two representations that correspond to the A-B or C-D relationship (with the intervening X-element abstracted away), the model can recognize novel test sequences that conform to those regularities. By contrast, when the model forms more interpretations, the middle X-elements are not abstracted away. In these cases, iMinerva is representing some information about specific X-elements that have occurred in middle position. This predicts that children would have more difficulty differentiating between rule-following and rule-violating test trials using novel combinations of A-X-B or C-X-D elements. Because children remember the particular A-X-B configurations they have seen, they have more difficulty generalizing to novel configurations, and thus look equivalently at rule-following or rule-violating test items.

In the 24 X-item case, the 24 intervening phoneme groups tend to be quickly abstracted away, and this caused the strong result in Figure 5 panel f. In the case of 24 intervening X elements, we see 65.0% of the simulations creating the two strong interpretations for the A-X-B and C-X-D intervening grammars, while with three intervening X-items we only have 31.2% creating two interpretations, and with 12 X-items only 28.8% of the simulations yield two interpretations. It is noteworthy that the interpretations in the 3 and 12 intervening items case were often weaker matches to the A-X-B and C-X-D patterns. This precision of the representations in the 24 intervening item case is made more clear in Figure 5 left by considering how even for a .75 threshold the 24 X-items shows a peak at 2, while the 3 and 12 conditions show a mix of interpretations, tending to peak around 6 (which indicates these simulations cannot differentiate the two grammars).

Insert Figure 6 about here

The power of our abstraction mechanism is revealed in Figure 6. Figure 6 looks at the interpretations stronger than 0.75 for each student in each condition, and then graphs the student average count of null abstracted features for the interpretations of each student. For example, if a student has three interpretations, with 5, 12, and 28 counts for null values, they would be represented by a value of 15 for their count of abstracted features. As we can see in Figure 6, abstraction is much poorer in the conditions with fewer intervening items. Furthermore, when abstraction is turned off, as shown in the Figure 7 result, the model is unable to abstract the intervening information, and fails to form any coherent representations except in a few exceptional cases. This ability of iMinerva to abstract differentiates it from MINERVA 2, which would be completely unable to produce results we have shown above because it does not produce abstractions in any way.

Insert Figure 7 about here

General Discussion

In recent years, a wide variety of experimental evidence has been advanced to support the claim that statistical learning plays a role in language development. While the original demonstrations of statistical learning focused on word segmentation (Hayes & Clark, 1970; Saffran et al., 1996), subsequent research has suggested that statistical

learning contributes to the discovery of syntactic structure (e.g. Hudson Kam & Newport, 2009; Thompson & Newport, 2007), word meaning (e.g. Vouloumanos & Werker, 2009), phonotactic patterns (e.g. Saffran & Thiessen, 2003), and phonemic categories (e.g. Maye et al., 2002). Each of these different aspects of language can be characterized, at least in part, by statistical regularities from which infants and adults are able to benefit. However, the regularities that contribute to learning in these domains are distributional: they relate to the frequency and variability of exemplars. Our goal in this series of simulations was to assess whether a single approach can explain learning from distributional regularities in all of these tasks.

The simulations of iMinerva represent one approach to assessing this possibility. Modeling provides an opportunity to determine whether a simple formal model is capable of handling a wide variety of learning problems. A demonstration that a single model *can* succeed at a variety of tasks does not conclusively prove that humans accomplish the task in a similarly uniform manner. Instead, modeling provides an existence proof that a unified approach is capable of succeeding at a wide variety of tasks, and allows for a conclusive demonstration of tasks in which the proposed learning mechanism fails. With this in mind, the iMinerva simulations presented in this paper suggest an important conclusion about the mechanisms underlying statistical learning. These simulations demonstrate that it is possible for a relatively simple memory-based approach to account for a wide variety of learning in tasks where the critical statistical feature is the distribution of information.

The three tasks that iMinerva simulated (a phonetic discrimination task, a word learning task, and a non-adjacent association learning task) are quite dissimilar on the

surface. Nevertheless, iMinerva learned successfully in each task. Note that this is not the only logical possibility; it might have been the case that iMinerva would be unable to simulate one or more of these tasks. Indeed, prior models of distributional learning have tended to focus primarily on only one of these problems (e.g., McMurray et al., 2009; Vallabha et al., 2007. Even those prior models that have been applied to multiple distributional learning problems have been constrained by the fact that their architectures have been focused on acquiring relatively domain-specific kinds of knowledge, such as a lexicon or a syntactic structure, meaning that they are not easily applied to other domains (e.g., Chang, Dell, & Bock, 2006; Feldman et al., 2009). The unique contribution of the iMinerva model is that its success in each of these three simulations indicates that it is possible to explain all three tasks in terms of domain general processes of long-term memory. In particular, the success of iMinerva suggests that learning in all three settings can be accounted for via the process of comparing between current and prior exemplars, and integrating them into a representation that is sensitive to the central tendencies of prior experience.

There are many other linguistically relevant learning tasks we did not simulate where the critical statistical feature is the distribution of events. For example, infants are able to learn phonotactic and phonological regularities from exposure to a set of words in which an acoustic feature (such as stress) consistently occurs in a particular position (e.g. Onishi, Chambers, & Fisher, 2002; Saffran & Thiessen, 2003; Thiessen & Saffran, 2007). Without simulating these learning tasks, it is impossible to confidently state that iMinerva would match human performance in identifying the distribution of acoustic features across word positions. However, it seems plausible that iMinerva's approach to learning

is a good fit for these tasks. Exposed to a set of words that follow a predominant pattern (such as word-initial stress), iMinerva's interpretations should converge on that pattern through the process of engagement. If the majority of word forms stored in memory have a particular acoustic pattern, the subsequent presentation of that pattern should activate prior instances and yield an interpretation that is consistent with that pattern.

Indeed, it is possible that the principles embodied in iMinerva are capable of explaining learning from distributional regularities generally, far beyond the three cases simulated here. While we have focused on those aspects of learning that have been proposed to play a role in language development, the processes invoked by iMinerva are domain-general principles that have been incorporated into many prior models of long-term memory. That is, if the processes invoked by iMinerva are responsible for learning from distributional information, the same kind of learning should be seen in many domains, because the processes in iMinerva are available in many domains. In fact, there is some evidence to suggest that some of the features of distributional learning seen for linguistic stimuli can also be seen for non-linguistic stimuli (e.g. Thiessen, 2011). However, like any model, iMinerva can only learn about those features that it encodes. It may well be the case that differences between linguistic and non-linguistic stimuli emerge as a function of differences in the salience and distribution of the features available in different domains.

Another reason that it is premature to claim that iMinerva is definitively capable of simulating all aspects of distributional learning is that each of the three tasks we have simulated can be characterized as receptive. In each task, the infant (and the model) are required to detect or perceive some regularity, but not to produce it. There are a variety

of distributional learning tasks in which learner's production is influenced by the distributional regularities to which they are exposed (e.g., Warker & Dell, 2006; Warker, Dell, Whalen, & Gereg, 2008). Currently, iMinerva does not attempt to model production, in much the same way that it does not attempt to model infants' behavioral responses. Instead, iMinerva is focused on identifying the representations underlying performance in learning tasks, and the processes that lead to the formation of those representations. To the extent that the same representations underlie performance in both receptive and productive tasks, iMinerva is potentially capable of simulating learning in both, but the current simulations do not assess this possibility.

While iMinerva can simulate distributional learning in a wide variety of tasks, it must be noted that some variation in the model's parameters is necessary to "fit" iMinerva to the different tasks. One possibility is that these parameter differences reflect developmental differences in the populations being modeled, as the Maye et al. (2002), Thiessen (2007), and Gomez (2002) experiments involved infants of different ages. An alternative possibility is that the different stimuli or procedures used in the tasks themselves cause changes in the way infants process the stimuli, changes that can be reflected by different parameter settings in iMinerva. This is an issue that can only be resolved by further experimentation and simulation. It may be especially helpful if further development of iMinerva enables it to make predictions in terms of behavioral responses like looking times, so that it will be possible to more closely assess the fit between iMinerva's responses and infant data. But while understanding the differences in parameters across different simulations is clearly important, it should not distract from the main finding of these simulations: it is possible to account for a wide range of

performance in statistical learning tasks that require learning from the distribution of events by using the same core processes. These simulations demonstrate that by storing prior exemplars, and integrating current and prior exemplars into an interpretation that embodies their central tendencies and variability, it is possible to account for a wide range of sensitivity to distributional information.

Relation to Conditional Statistical Learning

Despite the possibility of iMinerva simulating a wide variety of distributional learning, there are some kinds of statistical regularities that iMinerva – at least in its current form – is simply unable to detect. In particular, iMinerva (like the original Minerva model) has no way of segmenting input. Presented with a string of speech, iMinerva simply stores the string as a complete, unbroken vector. As such, iMinerva is unable to simulate word segmentation tasks, or any statistical learning task in which the end result of learning is to segment coherent units from a larger stimulus. These kinds of statistical learning tasks rely on *conditional* (as opposed to distributional) statistics, in that the units that are segmented are ones where the component elements (such as syllables within a word) have a strong conditional relationship (Aslin, Saffran, & Newport, 1998). Modeling these different statistical learning tasks allows for a comparison of the learning mechanisms that are necessary to take advantage of each kind of statistical information.

Successful models of conditional statistical learning tasks (like word segmentation) invoke processes where disparate elements of the input are associated into a single representation, consistent with evidence that infants and adults are extracting discrete representations from statistical learning tasks (e.g. Fiser & Aslin, 2005; Giroux

& Rey, 2009; Graf Estes, Evans, & Else-Quest, 2007). In chunking models, this association is the central mechanism of learning: syllables are linked together into larger chunks (Perruchet & Vinter, 1998). Similarly, in Bayesian models, segmentation is accomplished by testing hypotheses about which groupings of syllables are most likely to be words, and the end state of learning is a lexicon (e.g. Frank et al., 2010). Both of these approaches to modeling segmentation assume that the goal of the task is to discover words, discrete representations in which syllables have been extracted from the larger unit.

By contrast, iMinerva has no extraction process. Instead, iMinerva learns by forming an interpretation in which the current exemplar and the most active (above a similarity threshold) prior exemplar are merged into a new representation that converges toward their central tendency. It may be the case, then, that a mechanistic account of statistical learning requires two processes: one that extracts exemplars from fluent input, and one that integrates information across exemplars (for discussion, see Thiessen et al., under review). Conditional statistical learning tasks such as segmentation require the ability to extract, while distributional statistical learning requires the learner to compare across exemplars and identify their central tendency. If this characterization is correct, then the broader term statistical learning can be decomposed into (at least) two separate components. A challenge for future modeling will be to see if it is possible to incorporate sensitivity to both conditional and distributional statistical regularities within a single computational framework.

Incorporating extraction and integration into a single model is likely be necessary for a complete understanding of statistical learning, because these two processes

influence each other. For example, the output of conditional statistical learning (e.g., words) can serve as the input to subsequent distributional analysis (Thiessen & Saffran, 2003). When exposed to a set of words that follow a consistent phonotactic or phonological pattern, infants are able to learn that pattern (e.g. Saffran & Thiessen, 2003; Thiessen & Saffran, 2007). From our perspective, this can be explained as a distributional analysis. When exposed to a set of words, infants create an interpretation (in iMinerva's terminology) that accentuates their common features, and deemphasizes those aspects that are inconsistent. The words that infants extract from fluent speech via conditional statistical learning provide a set of word forms from which infants can discover distributional regularities.

Just as conditional statistical learning supports distributional statistical learning, distributional statistical learning can influence conditional statistical learning. When infants learn a distributional regularity, such as the relation between stress and word initial position in English, this knowledge affects their subsequent segmentation (Saffran & Thiessen, 2003; Thiessen & Saffran, 2007). For example, rather than segmenting a word with iambic stress, infants will group together syllables across word boundaries (such as *TARis* from *guiTAR is*) to maintain the expected relation between stress and word position (e.g. Johnson & Jusczyk, 2001). That is, the items that infants extract from fluent speech change as a function of the distributional regularities they have discovered. Theories and models of conditional statistical learning will benefit from attempting to incorporate these kinds of regularities into mechanistic accounts of segmentation (Perruchet & Tillmann, 2010). A complete model of statistical learning will need to be

able to account for both conditional and distributional learning, and the way that knowledge acquired from one form of statistical learning influences the other.

Summary

Distributional statistical learning refers to the ability to benefit from statistical features of the input such as the variability and frequency of exemplars. Recent research suggests that these distributional characteristics of the input play an important role in infant language development (e.g. Onishi et al., 2002; Thiessen & Saffran, 2007). Young language learners are able to take advantage of distributional information to identify several linguistic regularities, including phonemic categories (Maye et al., 2002), minimal pair distinctions (Thiessen, 2007), and simple syntactic patterns (Gomez, 2002). However, the very breadth of distributional learning raises an important question: can a single underlying mechanism achieve all of these different tasks?

As the iMinerva model demonstrates, it is indeed possible for a single domain-general approach, incorporating processes of long-term memory, to accomplish all three tasks. In the iMineva model, distributional learning occurs through a process of comparison and integration: the current exemplar is compared to prior exemplars (stored in memory), and then the current exemplar is integrated with the strongest prior exemplar (above a similarity threshold) to form an interpretation that accentuates their common features and de-emphasizes their contradictory features. We believe that this approach may, in fact, be capable of explaining many additional aspects of distributional learning in addition to the three simulated above. This is due to the fact that iMinerva relies on principles that are domain general and fundamental features of human memory. As iMinerva demonstrates, these basic properties may potentially account for many aspects

of language development, and suggests a deep connection between the mechanisms of human memory and the more recent statistical learning literature.

Appendix

Our model of the phenomenon in this paper is called interpretative Minerva (iMinerva) because it is an extension of Hintzman's MINERVA 2 model (Goldinger, 1998; Hintzman, 1988) with extended capabilities to maintain not only examples in memory, but also interpretations of examples. In the model, each new example that the learnerencounters is compared with prior examples to determine the similarity with these prior examples. This comparison is an automatic process of human cognition according to the model and corresponds to learner's basic ability to interpret new experience through the lens of old experience (learning is constructive). If multiple prior examples are similar, the learner selects the strongest of them. Assuming a prior example is selected; the learner engages with the prior example to modify it and create an interpretation. This synthetic interpretation is then recorded as a new memory item.

Each interpretation the learner forms functions like an example according to the model, so that once interpretations are formed, they are themselves engaged with by new examples. In this way, interpretations are like concepts that originally develop from a perceptual experience, but then get increasingly divorced from perceptual experiences as multiple perceptual encounters shape conceptual learning (Sloutsky, 2009). Learning in the model occurs as simple adjustment of the prior example or interpretation trace (or memory) using an additive learning rule to determine how prior memories grow by a proportion of the new similar example. This mechanism serves to create a more general representation from a class of items that originally share some similarity. If the new example has a feature that is different than an old example, despite being similar overall,

this learning averages the old feature with a portion of the new feature to reduce the strength of the feature that is different in the interpretation.

Interpretation in this way is a mechanism for prototype creation in the model. If, for example, prior example A and new example B are found to be similar, they may be engaged. In this case, if A has feature 1 = -1 and B has feature 1 = 1, then the interpretation created will show a feature 1 that is moved from -1 closer to 1 during learning. While this means that our interpretation is more general, that generality is specious (will not generalize) because feature 1 is still included in the interpretation. For this reason, if example A is repeated it will still match with prior example A better than the interpretation, while if example B is repeated it will also match with prior example B better than the interpretation.

To resolve this issue we introduce a very simple abstraction mechanism that removes features from interpretations when those features are some fraction of the maximum absolute feature strength. This abstraction mechanism seems a natural addition to the system, since salience is something that the brain seems to encode directly (Gottlieb, 2007) and our abstraction mechanism is inherently a mechanism that abstracts away less salient features. Not only do we find evidence for salience related information at the level of parietal cortex activity, but we also see that there appear to be cognitive benefits of using prototypes (Winkielman et al., 2006). Furthermore, there are very good reasons to believe that humans are limited in their focus of attention (Miller, 1956), and so, interpretations must necessarily become abstract because of the lack of an ability to attend to all the features in a stimulus each time it is encountered in the environment.

Specification

In iMinerva memory traces are represented as vectors of real numbers where some features may be null values. In contrast, MINERVA 2 requires 1, 0 or -1 values. This change in the feature coding provides a representation that allows us to capture both the strength and durability of an interpretation as the absolute value of a feature's strength. In this formalism, 0 comes to mean either a weak or very equivocal feature, and in either case, we allow features to transition to a null value when they are near 0. This mechanism (abstraction, described below) allows us to represent salience as a binary quantity that depends upon feature strength. This binary salience was a simplification of more complex alternatives that would have required feature salience as a continuous quantity for each feature. Table 3 shows how each syllable was coded with 16 features.

Insert Table 3 about here

Since we use this representation format for our interpretations, we also needed a new similarity function, as MINERA 2 simply uses the weighted average of feature agreement. Because our features now represent the strength of each feature in the interpretation, we are no longer looking for the mere agreement of features, but rather how the pattern of strengths in the exemplar is similar to the pattern of strengths in the interpretation. Because of this we have adopted a well-established measure, cosine similarity, because cosine similarity compares the magnitude pattern by including a normalization factor while MINERVA 2 similarity only weights yes or no agreement of features. In addition to handling the magnitudes, cosines similarity has a long history of use in text classification (Salton, 1989). Equation 1 below shows the cosine similarity

function. Like Hintzman's MINERVA 2, we compress the results of this metric to compute similarity by cubing the raw cosine similarities to increase dispersion amongst the values obtained. The calculation of cosine similarity is shown in Equation 1. Furthermore, we have modified this traditional equation such that if a feature is missing (has been abstracted away in the case of interpretations, or was not present in the stimuli in the case of examples) from either trace A or trace B, that feature is ignored in the computation.

$$\text{similarity}_{A,B} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

Each new example is compared with all prior memory traces (both interpretations and examples) to see if some similarity threshold parameter is exceeded. If the threshold is exceeded, it indicates that the learner notices the match(s) with prior stimuli. If nothing is matched, no interpretation is formed. While the underlying comparison process is assumed to unfold over time, the model operationalises the outcome by a simple "max similarity rule" Equation 2 shows how a prior trace, A, accumulates a portion of the strength of B when the similarity of B exceeds threshold and is the maximum similarity trace that exceeds threshold. The learning rate for this accumulation is represented by λ.

$$\begin{aligned} &C = \text{null} \\ &\text{if } S(A,B) < threshold \\ &\quad \text{or} \\ &C = A + \lambda B \\ &\text{if } S(A,B) < threshold \text{ and } B = max_T \text{ in all traces} S(A,T) \end{aligned} \quad (2)$$

This learning process is both strengthening (Equation 2) and abstractive. The abstractive component is captured in Equation 3 with specifies that given any feature in

the interpretation, it will be removed if it is weak relative to the maximal feature. This means that the absolute value of the strength of any feature must exceed a threshold for that feature to be retained. Equation 3 describes how each feature must exceed this criterion. Equation 3 uses the ρ parameter which is the fraction of the maximum absolute value that must be exceeded to retain a feature, otherwise that feature is set to null.

$$\begin{aligned} &\text{for features } i=1..n \\ &\quad C_i = \text{null} \\ &\quad \text{if } |C_i| < \rho(\max|C_{1..n}|) \end{aligned} \tag{3}$$

Finally, it seemed useful to provide some inclusion of forgetting in the model, which we simulate with simple exponential decay. While exponential decay may be a less accurate than power law or other functions in modeling forgetting (Rubin & Wenzel, 1996), in this model where forgetting is not a key factor, this decay mechanism adds plausibility to the model because it illustrates how memory traces are lost and why the model does not need to be concerned about the criticism that storing unlimited examples is implausible. The model is explicitly limited in the examples it can store because old examples eventually decay to the point they are never engaged, and are therefore essentially deleted. Equation 4 shows decay in the model for some example feature vector, N.

$$N_t = \delta N_{t-1} \tag{4}$$

Table 4 shows the parameters across the models, which are discussed in the paper body where appropriate. The model above is highly simplified to clarify explanation, but we argue that it captures the basic process of general exemplar learning and prototype extraction as it occurs in learners without complex language. We would not argue that the model above is correct or complete, merely that it adds to our understanding by showing a minimal set of principles that can achieve the patterns of representation we predict causes the behaviors in this paper. It seems likely that infants have minimal ways to direct the above processes, and that the cycle of experience, learning, and abstraction is driven by physical needs or by attraction to similarities in the environment (particularly similarities to items that were associated with reward in the past). No doubt, humans become quite skilled at guiding this cycle and sculpting their learning as their capabilities for action grow and they develop complex symbolic representations and goal structures.

Insert Table 4 about here

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of Conditional Probability Statistics by 8-Month-Old Infants. *Psychological Science, 9*(4), 321-324. doi: 10.1111/1467-9280.00063

Bomba, P. C., & Siqueland, E. R. (1983). The nature and structure of infant form categories. *Journal of Experimental Child Psychology, 35*(2), 294-328. doi: 10.1016/0022-0965(83)90085-1

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development, 10*(2), 159-199. doi: 10.1016/0885-2014(95)90008-x

Chomsky, N. (1959). A Review of B. F. Skinner's Verbal Behavior. *Language, 35*(1), 26-58. doi: citeulike-article-id:263746

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*(3), 804-809. doi: 10.1016/j.cognition.2008.04.004

Cohen, L. B. (1991). Infant attention: An information processing approach. In M. J. Weiss & P. R. Zalazo (Eds.), *Newborn attention: Biological constraints and the influence of experience* (pp. 1-21). Norwood, NJ: Ablex Publishing Corporation.

Cohen, L. B. (1998). An information-processing approach to infant perception and cognition. In F. Simion & G. Butterworth (Eds.), *The Development of Sensory, Motor, and Cognitive Capacities in Early Infancy* (pp. 277-300). East Sussex: Psychology Press.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 24-39. doi: 10.1016/s0278-7393(05)80004-7

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant Melodies: Statistical Learning of Nonadjacent Dependencies in Tone Sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(5), 1119-1130.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition, 105*(2), 247-299. doi: 10.1016/j.cognition.2006.09.010

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language, 60*(3), 351-367. doi: 10.1016/j.jml.2008.10.003

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can Infants Map Meaning to Newly Segmented Words? *Psychological Science, 18*(3), 254-260. doi: 10.1111/j.1467-9280.2007.01885.x

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National academy of Sciences of the United States of America, 99*(24), 15822-15826. doi: 10.1073/pnas.232472899

Fiser, J., & Aslin, R. N. (2005). Encoding Multielement Scenes: Statistical Learning of Visual Feature Hierarchies. *Journal of Experimental Psychology: General, 134*(4), 521-537. doi: 10.1037/0096-3445.134.4.521

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling Human Performance in Statistical Word Segmentation. *Cognition, 117*(2), 107-125.

Gibson, E. J. (1940). A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychological Review, 47*(3), 196-229. doi: : 10.1037/h0060582

Giroux, I., & Rey, A. (2009). Lexical and Sublexical Units in Speech Perception. *Cognitive Science, 33*(2), 260-272. doi: 10.1111/j.1551-6709.2009.01012.x

Goldinger, S. D. (1998). Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review, 105*(2), 251-279.

Gomez, R. L. (2002). Variability and Detection of Invariant Structure. *Psychological Science, 13*(5), 431-436. doi: 10.1111/1467-9280.00476

Gomez, R. L., & Maye, J. (2001). *Developmental trends in acquiring non-adjacent dependencies*. Unpublished manuscript, University of Arizona, Tucson.

Gottlieb, J. (2007). From Thought to Action: The Parietal Cortex as a Bridge between Perception, Action, and Cognition. [10.1016/j.neuron.2006.12.009]. *Neuron, 53*(1), 9-16.

Graf Estes, K., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the Nonword Repetition Performance of Children With and Without Specific Language

Impairment: A Meta-Analysis. *J Speech Lang Hear Res, 50*(1), 177-195. doi: 10.1044/1092-4388(2007/015)

Hall, G., & Honey, R. C. (1989). Contextual Effects in Conditioning, Latent Inhibition, and Habituation: Associative and Retrieval Functions of Contextual Cues. *Journal of Experimental Psychology: Animal Behavior Processes, 15*(3), 232-241.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition, 78*(3), B53-B64. doi: 10.1016/s0010-0277(00)00132-3

Hayes, J. R., & Clark, H. H. (1970). Experiments in the segmentation of an artificial speech analog. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 221-234). New York: Wiley.

Hintzman, D. L. (1984). Episodic versus semantic memory: A distinction whose time has come--and gone? *Behavioral and Brain Sciences, 7*(02), 240-241. doi: 10.1017/S0140525X00044435

Hintzman, D. L. (1986). "Schema Abstraction" in a Multiple-Trace Memory Model. *Psychological Review, 93*(4), 411-428. doi: 10.1037/0033-295x.93.4.411

Hintzman, D. L. (1988). Judgments of Frequency and Recognition Memory in a Multiple-Trace Memory Model. *Psychological Review, 95*(4), 528-551.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology, 59*(1), 30-66. doi: 10.1016/j.cogpsych.2009.01.001

Hunt, R. H., & Aslin, R. N. (2010). Category induction via distributional analysis: Evidence from a serial reaction time task. *Journal of Memory and Language, 62*(2), 98-112. doi: 10.1016/j.jml.2009.10.002

Johnson, E. K., & Jusczyk, P. W. (2001). Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics. *Journal of Memory and Language, 44*(4), 548-567. doi: 10.1006/jmla.2000.2755

Johnson, E. K., & Seidl, A. (2008). Clause Segmentation by 6-Month-Old Infants: A Crosslinguistic Perspective. *Infancy, 13*(5), 440-455.

Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics, 21*, 3-28.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition, 83*(2), B35-B42. doi: 10.1016/s0010-0277(02)00004-5

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., . . . Lacerda, F. (1997). Cross-Language Analysis of Phonetic Units in Language Addressed to Infants. *Science, 277*(5326), 684-686. doi: 10.1126/science.277.5326.684

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*(2), F13-F21. doi: 10.1111/j.1467-7687.2006.00468.x

Kuhl, P. K., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606-608. doi: 10.1126/science.1736364

Lidz, J., Gleitman, H., & Gleitman, L. (2003). Understanding how input matters: verb learning and the footprint of universal grammar. *Cognition, 87*(3), 151-178. doi: 10.1016/s0010-0277(02)00230-5

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant Rule Learning Facilitated by Speech. *Psychological Science, 18*(5), 387-391. doi: 10.1111/j.1467-9280.2007.01910.x

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule Learning by Seven-Month-Old Infants. *Science, 283*(5398), 77-80. doi: 10.1126/science.283.5398.77

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science, 11*(1), 122-134. doi: 10.1111/j.1467-7687.2007.00653.x

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101-B111. doi: 10.1016/s0010-0277(01)00157-3

McClelland, J. L., & Plaut, D. C. (1999). Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences, 3*(5), 166-168.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed Memory and the Representation of General and Specific Information. *Journal of Experimental Psychology: General, 114*(2), 159-188. doi: 10.1037/0096-3445.114.2.159

McMurray, B., Aslin, R.N., & Toscano, J.C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science, 12, 369-378.*

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition, 30*(5), 678-686. doi: 10.3758/bf03196424

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*(1), 91-117. doi: 10.1016/s0010-0277(03)00140-9

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). Sequential Expectations: The Role of Prediction-Based Learning in Language. *Topics in Cognitive Science, 2*(1), 138-153. doi: 10.1111/j.1756-8765.2009.01072.x

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*(2), 127-162. doi: 10.1016/s0010-0285(03)00128-2

Onishi, K. H., Chambers, K. E., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition, 83*(1), B13-B23. doi: 10.1016/s0010-0277(01)00165-2

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences, 105*(7), 2745-2750. doi: 10.1073/pnas.0708424105

Pater, J., Stager, C., & Werker, J. F. (2004). The lexical acquisition of phonological contrasts. *Language, 80*(3), 361-379.

Perruchet, P., & Tillmann, B. (2010). Exploiting Multiple Sources of Information in Learning an Artificial Language: Human Data and Modeling. *Cognitive Science, 34*(2), 255-285. doi: 10.1111/j.1551-6709.2009.01074.x

Perruchet, P., Tyler, M. D., Galland, N., & Peereman, R. (2004). Learning Nonadjacent Dependencies: No Need for Algebraic-Like Computations. *Journal of Experimental Psychology: General, 133*(4), 573-583.

Perruchet, P., & Vinter, A. (1998). Feature creation as a byproduct of attentional processing. *Behavioral and Brain Sciences, 21*(01), 33-34.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(6), 906-914. doi: 10.1002/wcs.78

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*(4), 734-760.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology, 39*(3), 484-494. doi: citeulike-article-id:933733

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*: Addison-Wesley Longman Publishing Co., Inc. .

Seidenberg, M. S., & McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review, 96*(4), 523-568.

Shvachkin, N. K. (1973). The development of phonemic speech perception in early childhood. In C. Ferguson & D. Slobin (Eds.), *Studies of Child Language*.

Sloutsky, V. M. (2009). From Perceptual Categories to Concepts: What Develops?

Shi, L., Griffiths, T.L., Feldman, N.H., & Sanborn, A.N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychological Bulletin and Review, 17*, 443-464.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*(6640), 381-382.

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language, 56*(1), 16-34. doi: 10.1016/j.jml.2006.07.002

Thiessen, E. D. (2009). Statistical learning. In E. Bavin (Ed.), *Cambridge Handbook of Child Language* (pp. 35-50). Cambridge: Cambridge University Press.

Thiessen, E. D. (2011). Domain General Constraints on Statistical Learning. *Child Development, 82*(2), 462-470. doi: 10.1111/j.1467-8624.2010.01522.x

Thiessen, E. D., Kronstein, & Hufnagle. (under review). The extraction and integration framework: A two-process account of statistical learning.

Thiessen, E. D., & Saffran, J. R. (2003). When Cues Collide: Use of Stress and Statistical Cues to Word Boundaries by 7- to 9-Month-Old Infants. *Developmental Psychology, 39*(4), 706-716.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to Learn: Infants' Acquisition of Stress-Based Strategies for Word Segmentation. *Language Learning and Development, 3*(1), 73-100. doi: 10.1207/s15473341lld0301_3

Thiessen, E. D., & Yee, M. N. (2010). Dogs, Bogs, Labs, and Lads: What Phonemic Generalizations Indicate About the Nature of Children's Early Word-Form

Representations. *Child Development, 81*(4), 1287-1303. doi: 10.1111/j.1467-8624.2010.01468.x

Thompson, S. P., & Newport, E. L. (2007). Statistical Learning of Syntax: The Role of Transitional Probability. *Language Learning and Development, 3*(1), 1 - 42.

Toro, J. M., Nespor, M., Mehler, J., & Bonatti, L. L. (2008). Finding Words and Rules in a Speech Stream. *Psychological Science, 19*(2), 137-144. doi: 10.1111/j.1467-9280.2008.02059.x

Toro, J. M., & Trobalon, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception & psychophysics, 67*(5), 867-875.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104*(33), 13273-13278. doi: 10.1073/pnas.0705369104

Van Rijn, H., & Anderson, J. (2003). *Modeling lexical decision as ordinary retrieval*.

Vouloumanos, A., & Werker, J. F. (2009). Infants' Learning of Novel Words in a Stochastic Environment. *Developmental Psychology, 45*(6), 1611-1617. doi: 10.1037/a0016134

Warker, J. A., & Dell, G. S. (2006). Speech Errors Reflect Newly Learned Phonotactic Constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(2), 387-398. doi: 10.1037/0278-7393.32.2.387

Warker, J. A., Dell, G. S., Whalen, C. A., & Gereg, S. (2008). Limits on Learning Phonotactic Constraints From Recent Production Experience. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 34(5), 1289-1295. doi: 10.1037/a0013033

Werker, J. F., & Curtin, S. (2005). PRIMIR: A Developmental Framework of Infant Speech Processing. *Language Learning and Development, 1*(2), 197-234. doi: citeulike-article-id:3937582

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' Ability to Learn Phonetically Similar Words: Effects of Age and Vocabulary Size. *Infancy, 3*(1), 1-30. doi: 10.1207/s15327078in0301_1

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*(1), 49-63. doi: 10.1016/s0163-6383(84)80022-3

Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes Are Attractive Because They Are Easy on the Mind. *Psychological Science, 17*(9), 799-806.

Table 1. Feature coding for the stimuli used in Simulation 1. The first feature represents voicing, on a continuum from 1 (voiced) to -1 (voiceless). The other features specify consonant and vowel features such as place of articulation and vowel height.

	Features																Bimodal	Unimodal
Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Frequency	Frequency
da4	1	1	0	0	1	0	1	-1	-1	-1	0	0	0	0	0	0	4	4
da3	0.715	1	0	0	1	0	1	-1	-1	-1	0	0	0	0	0	0	16	4
da2	0.43	1	0	0	1	0	1	-1	-1	-1	0	0	0	0	0	0	8	8
da1	0.145	1	0	0	1	0	1	-1	-1	-1	0	0	0	0	0	0	4	16
ta1	-0.145	1	0	0	1	0	1	-1	-1	-1	0	0	0	0	0	0	4	16
ta2	-0.43	1	0	0	1	0	1	-1	-1	-1	0	0	0	0	0	0	8	8
ta3	-0.715	1	0	0	1	0	1	-1	-1	-1	0	0	0	0	0	0	16	4
ta4	-1	1	0	0	1	0	1	-1	-1	-1	0	0	0	0	0	0	4	4

Table 2. Vectors for suffixes in Thiessen (2007).

	Features															
Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
-bow	1	1	1	0	0	0	-1	1	-1	1	0	0	0	0	0	0
-goo	1	1	0	0	0	1	-1	1	1	1	0	0	0	0	0	0

Table 3. Descriptions of the feature meanings for the phoneme vectors used by iMinerva.

Feature Index	Feature Description
1	Initial consonant - voiced/voiceless
2	Initial consonant - stop/non-stop
3	Initial consonant - labial
4	Initial consonant - dental
5	Initial consonant – alveolar
6	Initial consonant – glottal
7	Vowel - front/back
8	Vowel - high/low
9	Vowel - rounded/unrounded
10	Vowel - long/short
11	Final consonant - voiced/voiceless
12	Final consonant - stop/non-stop
13	Final consonant - labial
14	Final consonant - dental
15	Final consonant - alveolar
16	Final consonant - glottal

Table 4. Parameter values used in the simulations.

Parameter	Represents	Maye	Thiessen	Gomez
λ	learning rate	0.2	0.2	0.2
ρ	abstraction proportion	0.1	0.1	0.6
δ	decay rate	0.98	0.98	0.98
threshold	engagement threshold	0.85	0.6	0.45
threshold noise	*SD* of threshold (for each learner)	0.025	0.05	0.05

Figure Captions

Figure 1. This figure illustrates the growth of the /t/ and /d/ feature across examples and interpretations. Each step of the index indicates an example trace or interpretation is added to memory. The top figure shows bimodal input, while the bottom figure unimodal input.

Figure 2. Percentage of children with each interpretation count with a mean feature strength of 2.5 or greater after each of the 500 simulations for each of the two conditions.

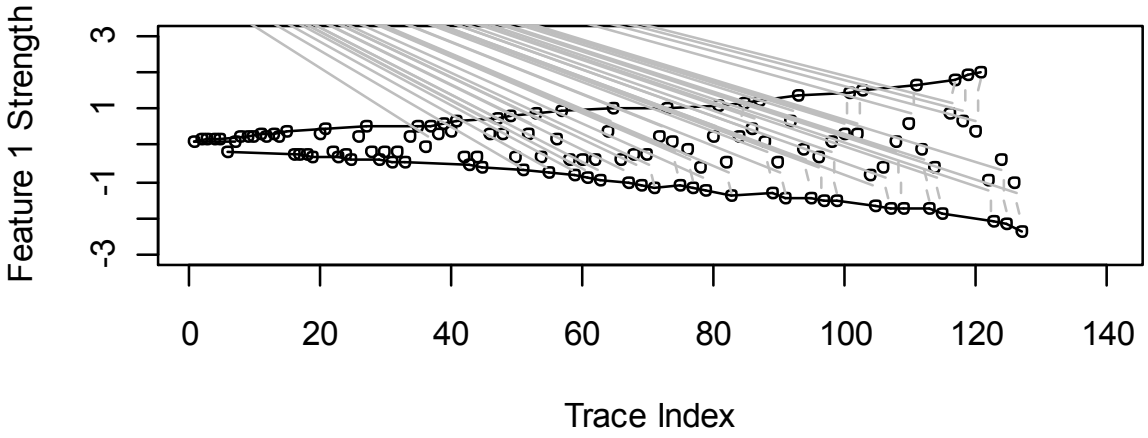
Figure 3. Percentage of simulations that form one, two, three or four interpretations with a mean feature strength of 2.5 or greater after the 500 simulations in the identical contexts and distinct contexts conditions.

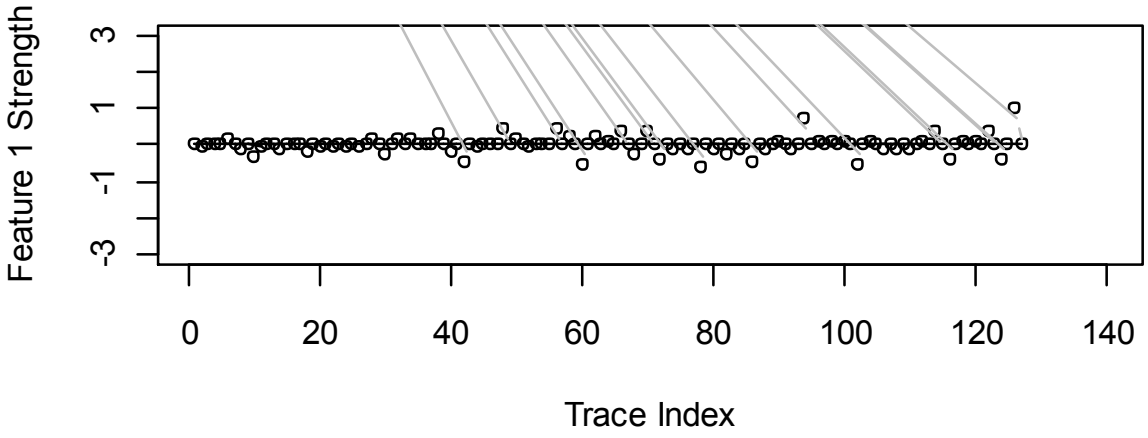
Figure 4. Examples of the interpretation that results a) from the presentation of either daw or taw in the daw, dawgoo, tawgoo case, b) from the presentation of daw in the daw, dawbow and tawgoo case, c) from the presentation of taw in the daw, dawbow and tawgoo case

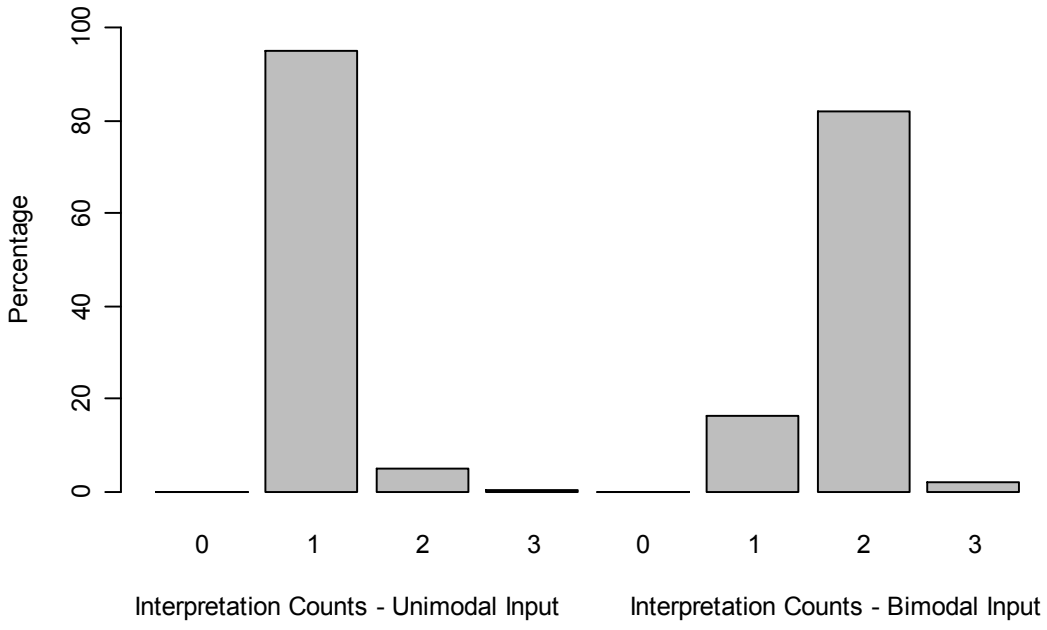
Figure 5. Percentage of simulated children with each interpretation count with a mean feature strength of 0.75 or greater (on the left; figures a, c, and e) or 2.0 or greater (on the right; figures b, d, and f) from 500 simulated learners for each simulation.

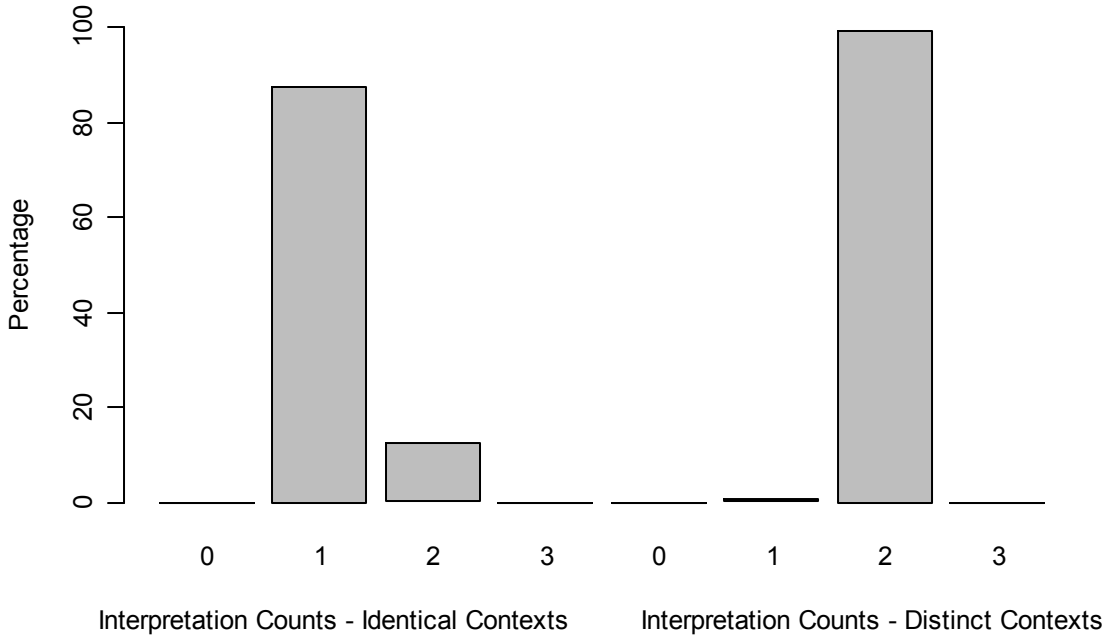
Figure 6. Average student abstraction in final interpretations with average feature strength greater than 0.75. Histograms show the count of abstractions for the 32 intervening syllables in each simulation for 3, 12, or 24 intervening items.

Figure 7. Unique interpretations with a mean feature strength of 2.0 or greater from 500 simulated learners for the 24 intervening item condition with abstraction set equal to 0.1 (from Simulations 1 and 2) compared to 0.6 in Figure 5 part f.

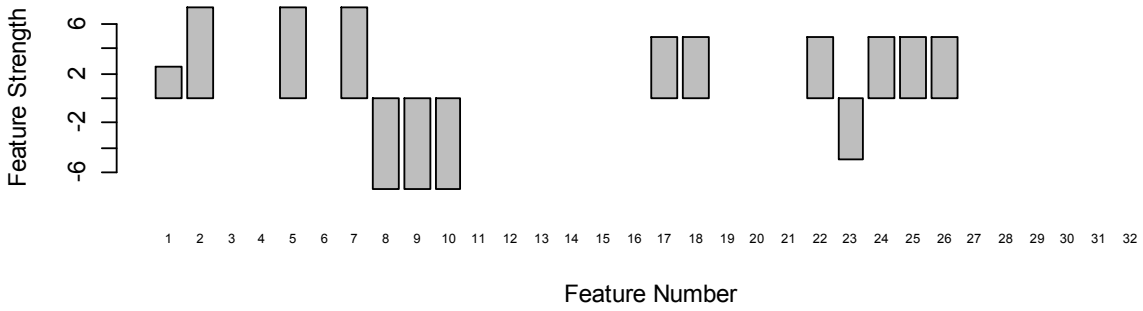

Feature 1 Strength
3
1
-1
-3
0
20
40
60
80
100
120
140
Trace Index


Feature 1 Strength
3
1
-1
-3
0
20
40
60
80
100
120
140
Trace Index

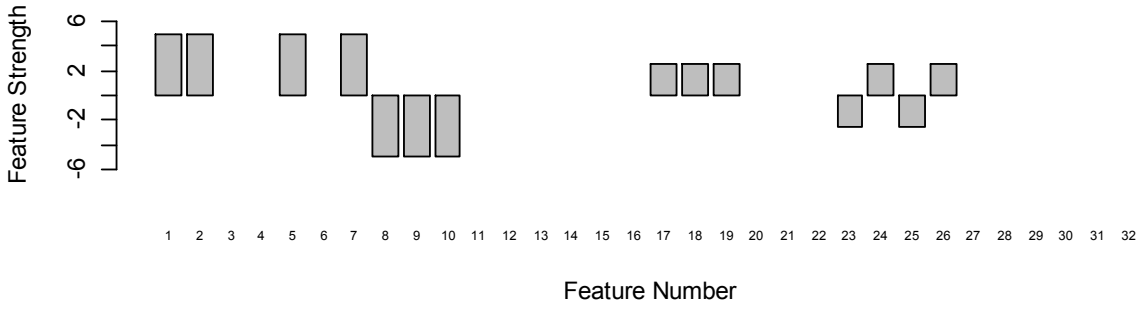

Percentage
0
20
40
60
80
100
0
1
2
3
0
1
2
3
Interpretation Counts - Unimodal Input
Interpretation Counts - Bimodal Input


Percentage
0
20
40
60
80
100
0
1
2
3
0
1
2
3
Interpretation Counts - Identical Contexts
Interpretation Counts - Distinct Contexts

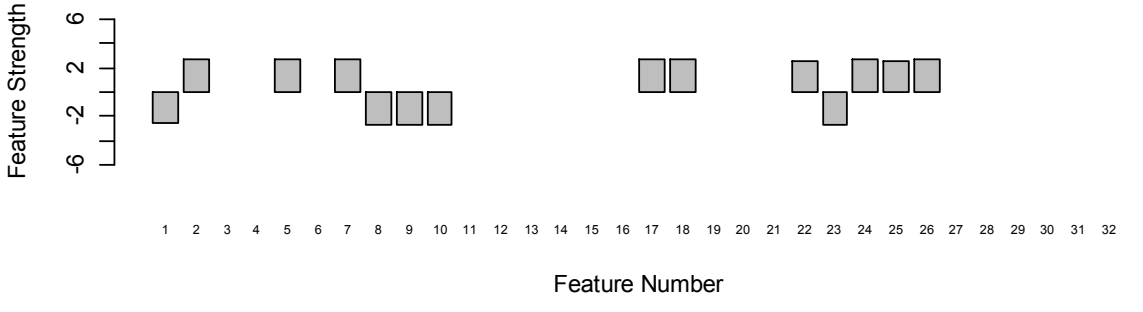
a)

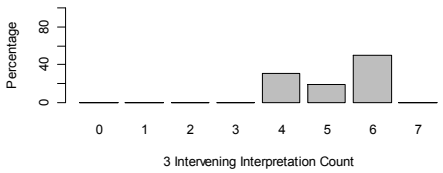

Feature Strength
6
2
-2
-6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Feature Number

b)

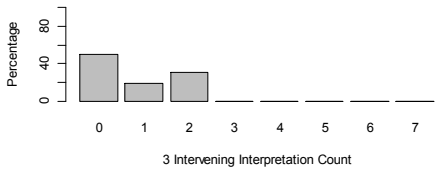

Feature Strength
6
2
-2
-6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Feature Number

c)

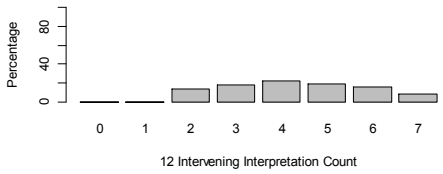

Feature Strength
6
2
-2
-6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Feature Number


Percentage
0
40
80
0 1 2 3 4 5 6 7
3 Intervening Interpretation Count

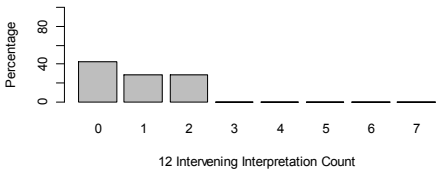
a)


Percentage
0
40
80
0 1 2 3 4 5 6 7
3 Intervening Interpretation Count

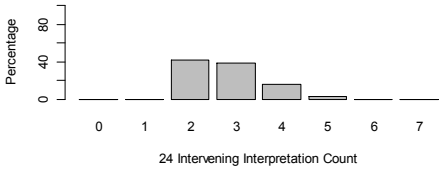
b)


Percentage
0
40
80
0 1 2 3 4 5 6 7
12 Intervening Interpretation Count

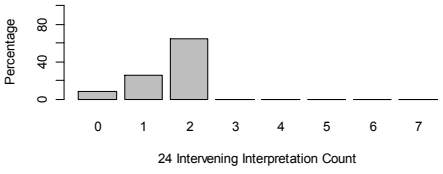
c)


Percentage
0
40
80
0 1 2 3 4 5 6 7
12 Intervening Interpretation Count

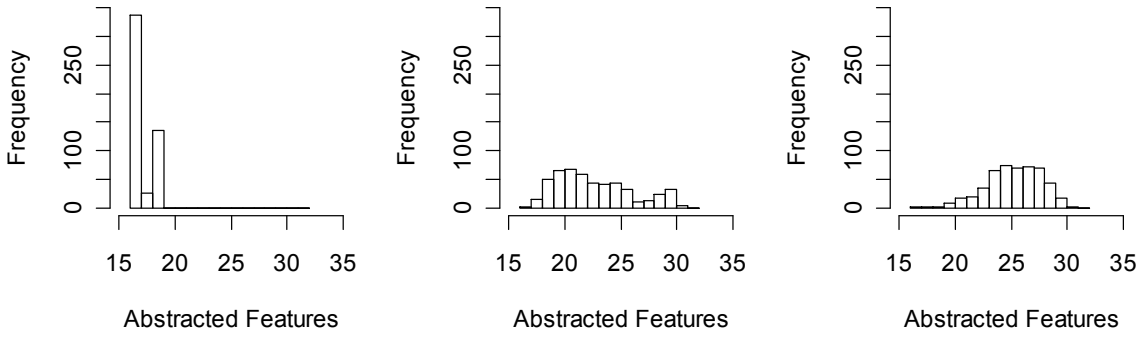
d)


Percentage
0
40
80
0 1 2 3 4 5 6 7
24 Intervening Interpretation Count

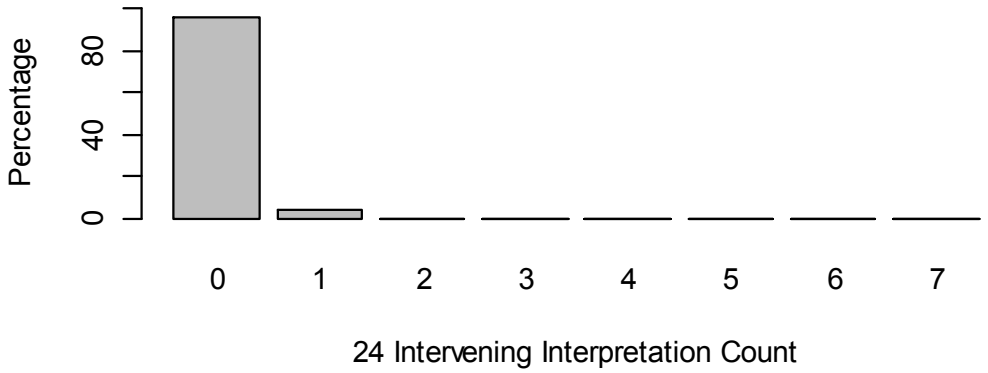
e)


Percentage
0
40
80
0 1 2 3 4 5 6 7
24 Intervening Interpretation Count

f)



a) 3 intervening b) 12 intervening c) 24 intervening


Percentage
0
40
80
0
1
2
3
4
5
6
7
24 Intervening Interpretation Count