Psychological Bulletin
2013, Vol. 139, No. 4, 792–814

0033-2909/13/$12.00 DOI: 10.1037/a0030801

The Extraction and Integration Framework: A Two-Process Account of Statistical Learning

Erik D. Thiessen, Alexandra T. Kronstein, and Daniel G. Hufnagle
Carnegie Mellon University

The term *statistical learning* in infancy research originally referred to sensitivity to transitional probabilities. Subsequent research has demonstrated that statistical learning contributes to infant development in a wide array of domains. The range of statistical learning phenomena necessitates a broader view of the processes underlying statistical learning. Learners are sensitive to a much wider range of statistical information than the conditional relations indexed by transitional probabilities, including distributional and cue-based statistics. We propose a novel framework that unifies learning about all of these kinds of statistical structure. From our perspective, learning about conditional relations outputs discrete representations (such as words). Integration across these discrete representations yields sensitivity to cues and distributional information. To achieve sensitivity to all of these kinds of statistical structure, our framework combines processes that extract segments of the input with processes that compare across these extracted items. In this framework, the items extracted from the input serve as exemplars in long-term memory. The similarity structure of those exemplars in long-term memory leads to the discovery of cues and categorical structure, which guides subsequent extraction. The extraction and integration framework provides a way to explain sensitivity to both conditional statistical structure (such as transitional probabilities) and distributional statistical structure (such as item frequency and variability), and also a framework for thinking about how these different aspects of statistical learning influence each other.

Keywords: statistical learning, language development, implicit learning, word learning

Humans live in a world filled with statistical regularities. Balls thrown into the air typically fall back to earth; nouns such as *dog* or *boy* are typically preceded by articles such as *a* or *the*. There is no doubt that learners are sensitive to these statistical regularities. One term to describe the ability to detect and use statistical structure is *statistical learning*. Saffran, Aslin, and Newport (1996) proposed this term to describe infants' ability to identify word boundaries solely from the statistical relation between sounds in the input. It is now widely acknowledged that infants and adults encode the statistical structure of their environment in a variety of tasks, including sequence learning (e.g., Haith, Wentworth, & Canfield, 1993; Stadler, 1992), category boundary detection (e.g., Maye, Werker, & Gerken, 2002), word-object association (Smith & Yu, 2008), cue-category association (Thiessen & Saffran, 2007), and causal learning (Sobel & Kirkham, 2007). Statistical learning likely plays a role in many different aspects of development, but it is thought to play an especially crucial role in language development. The discovery that infants are capable of benefiting from statistical structure in the input led to a reevaluation of the role of learning in language acquisition, after several decades of theoretical claims that learning played a relatively minor role in the process (e.g., Chomsky, 1980; Lidz, Gleitman, & Gleitman, 2003).

But the discovery of the importance of statistical learning has, in turn, raised a new set of issues. Perhaps the most important of these is the need for a definition of the mechanism (or mechanisms) that makes statistical learning possible. Consider the breadth of learning phenomena to which the term *statistical learning* is applied. One example of statistical learning is identifying conditional relations among elements of the input, allowing learners to detect that some aspects of the input are likely to predict each other or "go together," like sounds within a word (e.g., Aslin, Saffran, & Newport, 1998). Infants and adults are sensitive to conditional relations in both sequentially (e.g., auditory) and simultaneously (e.g., visual) presented stimuli—but it is not clear whether learning from both kinds of input is accomplished by the same or different mechanisms (e.g., Conway & Christiansen, 2006; Kirkham, Slemmer, & Johnson, 2002). And even beyond conditional relations, recent work has established that there are a variety of other statistical relations that influence learning. For example, infants are able to discover category boundaries simply from exposure to a distribution of exemplars differing in frequency (e.g., Maye et al., 2002). In addition, infants and adults are sensitive to the correlation between perceptual features and aspects of the input that are not directly perceptible (such as word boundaries) and learn to use these perceptual features as cues (e.g., Thiessen & Saffran, 2003).

This article was published Online First December 10, 2012.

Erik D. Thiessen, Alexandra T. Kronstein, and Daniel G. Hufnagle, Department of Psychology, Carnegie Mellon University.

Preparation of this article was funded in part by National Science Foundation Grant BCS-0642415 awarded to Erik D. Thiessen. We thank David Rakison and David Plaut for the willingness to read early (and much longer) drafts of the article.

Correspondence concerning this article should be addressed to Erik D. Thiessen, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: thiessen@andrew.cmu.edu

As this brief discussion indicates, humans are sensitive to both statistical information contained in a single stimulus (e.g., a word) and statistical information that can be integrated across several different stimuli (e.g., the exemplars composing a category). The breadth of statistical learning phenomena presents a challenge for a mechanistic account of statistical learning. It is not clear whether the same mechanism is responsible for sensitivity to both kinds of statistical information. Furthermore, it is not clear how these kinds of information interact, and when learners would favor one kind of information over another (e.g., E. K. Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003). There have been several mechanisms proposed to explain statistical learning. However, the vast majority of these mechanistic accounts have been solely focused on sensitivity to conditional relations, primarily in the context of word segmentation (e.g., Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Perruchet & Vinter, 1998). These models typically ignore, or are unable to account for, sensitivity to statistical information that requires integrating information across exemplars. Similarly, models of information integration are not intended to account for sensitivity to conditional relations, or segmentation of fluent input into smaller chunks (e.g., Feldman, Griffiths, & Morgan, 2009; Hintzman, 1984). That is, most prior theoretical accounts of statistical learning are single-process accounts: They focus on one aspect of statistical learning and are agnostic about other aspects, or how different aspects of statistical learning might interact.

Our goal in this review was to develop a framework to explain statistical learning that is able to incorporate the multiple different kinds of statistical relations to which learners are sensitive. The framework we propose accounts for statistical learning in terms of processes that are an integral part of memory. That is, we believe that sensitivity to statistical information in the environment arises directly from the way that humans (and other species) store and access information in memory. In particular, we invoke two memory-based processes to explain statistical learning: *extraction* and storage of statistically coherent units (such as words) from the input and *integration* across units stored in memory to identify central tendency. This account assumes that learning involves the extraction and storage of statistically coherent chunks of the input (for evidence in favor of this assumption, see the Discussion in the the *Extraction and Conditional Statistics section). Storing these statistically coherent clusters provides an opportunity for integration of information across the items in memory. Processes of integration are routinely invoked in models of long-term memory, especially those concerned with sensitivity to prototypes and category formation (e.g., Hintzman, 1984; McClelland & Rumelhart, 1985). Integration will also allow learners to detect those features that are common across prior exemplars, an ability that has been argued to play an important role in cue learning (e.g., Thiessen & Saffran, 2007).

The framework we are advocating extends beyond prior models of statistical learning, in that it encompasses more than just learning of conditional relations. By doing so, it provides an opportunity to explore the interplay between different kinds of statistical information in the environment. This has been an important topic of research (e.g., E. K. Johnson & Jusczyk, 2001; E. K. Johnson & Seidl, 2008; Saffran & Thiessen, 2003), but one that has not been systematically incorporated into theoretical accounts of statistical learning (though see Adriaans & Kager, 2010). Additionally, this framework provides an opportunity to generate hypotheses about the development of statistical learning. Much prior research has sought to explain how statistical learning might contribute to development (e.g., Maye et al., 2002; Saffran et al., 1996; Thiessen & Saffran, 2003). Much less is known about how statistical learning itself develops and changes with age (though see Hudson Kam & Newport, 2009; Kirkham et al., 2002). By advancing an account of the processes underlying statistical learning, it becomes possible to make principled predictions about how the operation of those processes should change with age, which provides an opportunity to synthesize research on statistical learning with a broader developmental view of the factors that should influence statistical learning.

As such, the goal of this review was, first, to provide an overview of the characteristics of statistical learning. We suggest that the term *statistical learning* refers to three qualitatively different kinds of tasks: tasks that require sensitivity to conditional relations (e.g., word segmentation), tasks that require sensitivity to distributional information (e.g., category learning), and tasks that require sensitivity to the relation between perceptual characteristics of the input and the units that organize the input (e.g., cue learning). Second, we provide a framework that accounts for all of these disparate statistical learning phenomena. This framework combines processes responsible for extracting units (e.g., words or shapes) from the input with processes responsible for integrating information across those units, and can therefore provide an account for sensitivity to both statistical information relevant to a particular stimulus (accomplished via extraction) and arising from a comparison across multiple stimuli (accomplished via integration).

In the first section of this review, we discuss the kinds of statistical structures to which humans are sensitive and propose that these can be grouped into three categories: conditional, distributional, and cue-based statistics. As we discuss in the second section, extraction provides an account for sensitivity to conditional statistical information (such as that used in word segmentation), but is insensitive to other forms of statistical information. Conversely, as we discuss in the third section, integration across exemplars provides an account for sensitivity to distributional information, but no explanation for how exemplars are initially segmented from the input. But a framework combining these processes, as we discuss in the fourth section, is capable of accounting for the full range of statistical learning phenomena, including detection of category boundaries (e.g., Maye et al., 2002; Vallabha, McClelland, Pons, Werker, & Amano, 2007), and the discovery of useful cues for subsequent learning (e.g., Rakison & Lupyan, 2008; Thiessen & Saffran, 2003), and is also capable of explaining how different kinds of statistical information interact and influence each other both in a single set of stimuli (e.g., Saffran & Thiessen, 2003) and across developmental time (e.g., Thiessen & Saffran, 2003).

Defining Statistical Learning: What Statistical Relations Are Learned?

Any successful account of statistical learning must be rooted in a consideration of the kinds of statistical relations to which learners are sensitive. The term statistical learning has been applied to a wide variety of situations in which learners demonstrate some sensitivity to the statistical structure of the input. Though all of

these learning feats have been grouped under the same umbrella term *statistical learning*, no unified mechanistic account has been set forth to account for all of them. The vast majority of modeling work on statistical learning has focused on sensitivity to conditional relations, such as those used to segment words (e.g., Christiansen, Allen, & Seidenberg, 1998; Frank et al., 2010). But adults, infants, and animals are sensitive to a variety of other statistics that are not easily captured in terms of conditional relations. For descriptive purposes, we separate statistical learning into three categories: conditional, distributional, and cue-based statistical learning.

Conditional Statistics

Conditional statistics measure the predictive relationship between two events X and Y. Transitional probability (Harris, 1955; Hayes & Clark, 1970; Saffran et al., 1996) is a commonly used example of a conditional statistic and describes the likelihood that event Y will occur given information that some other event X has occurred. When X regularly predicts Y, transitional probabilities are high; transitional probability is low when X rarely predicts Y. For example, if X occurs 100 times, and the sequence X-Y occurs 30 times, the transitional probability between X and Y is 30%. Conditional statistics are a more robust metric of the strength of the relation between two events than the simple frequency of their co-occurrence (e.g., Aslin et al., 1998). This is due to the fact that two items can occur together quite frequently simply by virtue of the fact that they are both high-frequency items in the input. For example, a phrase like "the man" is relatively common, because both words are high frequency. However, because "the" can be followed by many other words, the transitional probability between "the" and "man" is low.

Human infants and adults are sensitive to conditional statistics, as are a variety of nonhuman animal species (Aslin et al., 1998; Toro & Trobalón, 2005). This sensitivity has been investigated most closely in the context of word segmentation. Across languages, sounds within a word are more predictable than sounds across word boundaries (e.g., Harris, 1955). For example, *copter* is very likely to come next after *heli*, but many different words could occur after *happy*. To explore whether learners are sensitive to this statistical property, experimenters have used artificial languages containing no cues to word segmentation except the statistical relations among sounds within and across word boundaries. Saffran et al.'s (1996) experiments provide a paradigmatic example of this experimental strategy. Infants were exposed to fluent speech containing four nonsense words: *pabiku*, *golabu*, *padoti*, and *tudaro*. These words occurred in random order, with no pauses between words, for 90 repetitions apiece (approximately 2.5 min). In the test phase, infants were presented with words, and their looking time to words was compared with one of two different kinds of foil items: nonwords (syllables that never occurred together in the speech stream, such as *kutiro*) or part-words (syllables that occurred across word boundaries, such as *bupado*). Infants' looking times revealed that they were able to discriminate between words and nonwords, and also between words and part-words. Statistical coherence is high in word test items, and low in nonwords and part-words. The finding that a discrimination between highly coherent and less coherent test items is quickly learned has been replicated many times, indicating that learners are sensitive to statistical coherence in the speech stream (e.g., Aslin et al., 1998; E. K. Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003).

Sensitivity to conditional statistical relations is not limited to relations between sound sequences within a speech stream. Learners also display sensitivity to sequential conditional statistics in tactile, visual, and tonal stimuli (Conway & Christiansen, 2006; Kirkham et al., 2002; Saffran, Johnson, Aslin, & Newport, 1999). Furthermore, learners are sensitive to conditional relations among sequences of actions, both their own and others' (e.g., Baldwin, Andersson, Saffran, & Meyer, 2008; Stadler, 1992). Although most demonstrations of sensitivity to conditional relations have occurred with sequential presentation, humans are also sensitive to conditional relations among elements presented simultaneously (e.g., Fiser & Aslin, 2005). This occurs when participants are shown complex scenes made up of several individual shapes. Across these scenes, the presence of one shape predicts the co-presence of a second shape, or multiple additional shapes. At test, learners discriminate between groupings of shapes that consistently occurred together and groupings of shapes that were less likely to co-occur (e.g., Fiser & Aslin, 2002). Note that in these laboratory demonstrations of sensitivity to conditional statistical relations, test trials typically involve discrimination between sequences with very high conditional relations (often approaching, or equal to, 100%) and sequences with far lower conditional relations. It is as yet unclear what minimum difference in likelihood is necessary to differentiate between conditional relations with different strength.

Critically, human sensitivity to statistical relations is not limited to elements that occur adjacently. Many of the relations that infants and adults learn involve regularities between elements that are not immediately adjacent, as is often the case in language. Although *the* predicts a subsequent noun, the noun can follow several words later (as in *the* surly *professor*). If statistical sensitivity were limited to detecting relations between adjacent items, statistical learning would be a severely limited learning tool. This is not the case. Several experiments have demonstrated that infant and adult learners detect conditional relations between nonadjacent elements, both in linguistic and nonlinguistic stimuli (Creel, Newport, & Aslin, 2004; Newport & Aslin, 2004). In sequences where X and Y are separated by intervening, unpredictable elements—such that the input might include sequences like XAY, XBY, and XCY—learners detect that X predicts that Y will follow. This kind of nonadjacent learning appears to be more difficult than detecting adjacent relations. For both audio and visual stimuli, the detection of nonadjacent relationships must be supported by high variability among intervening elements (Gómez, 2002) or by a perceptual cue linking the nonadjacent elements (e.g., Baker, Olson, & Behrmann, 2004; Creel et al., 2004). This kind of support is not necessary for the discovery of adjacent conditional relations.

Distributional Statistics

Conditional statistics reflect the strength of the relationship between elements X and Y. In contrast, distributional statistics reflect the central tendency, or prototypical characteristics, of a set of elements. These statistics are referred to as *distributional* because learners are sensitive to the frequency and variability of exemplars in the input (e.g., Maye, Weiss, & Aslin, 2008; Mintz,

2006; Thiessen, 2009). Maye et al.'s (2002) experiments on the effect of distribution of phonemic exemplars on infants' discrimination provide a paradigmatic example of distributional statistical learning. Maye et al. found that infants' categorical boundaries for phonetic distinctions, such as /d/ and (unaspirated) /t/, could be shifted on the basis of the frequency with which they were exposed briefly to exemplars. When exposed to a bimodal distribution of sounds, such that a prototypical /d/ and a prototypical unaspirated /t/ occurred frequently, infants were more likely to discriminate between exemplars of the two categories. When exposed to a unimodal distribution, whereby a sound intermediate between /d/ and unaspirated /t/ occurred most frequently, infants were less likely to discriminate between exemplars from the two categories even though these exemplars occurred equally often in the two training regimes. Subsequent work has replicated this result for other phonetic categories (e.g., Maye et al., 2008).

Sensitivity to the frequency of phonemic exemplars may explain how infants adapt to the phonemic structure of their native language in the first year of life (Werker & Tees, 1984). Sounds near the prototypical center of a phonemic category that a language uses occur more frequently than sounds at the boundaries between such categories (Werker et al., 2007). The role of distributional statistics in language acquisition is not limited to perception, however. The frequency of exemplars in the input has been suggested to play an important role in word learning and discovering syntactic patterns (e.g., Reber & Lewis, 1977; Thiessen & Yee, 2010). Like sensitivity to conditional information, sensitivity to distributional information plays a role in many domains in addition to language. Learners show sensitivity to statistics relating to categorical structure in domains, including visual perception (e.g., Dougherty & Haith, 2002), auditory perception (e.g., Lotto, Kluender, & Holt, 1997), and object categorization (e.g., Rakison, 2004; Younger & Cohen, 1986). Additionally, many nonhuman animal species are sensitive to distributional statistics (e.g., Lotto et al., 1997), which suggests that learning from distributional statistics is a domain-general ability rather than a language-specific one.

The second feature of the distributional structure of the input to which learners are sensitive is variability. When exposed to distributions with high variability, learners accept a wider range of exemplars as members of a category and are correspondingly less certain when required to make judgments about stimuli near a category boundary (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008). When there is very low variability in the input, category boundaries are comparatively sharper. Note that this discussion of variability presumes that learners are able to detect within-category variation, which is not represented according to classical theories of categorical perception (e.g., Liberman, Harris, Hoffman, & Griffith, 1957). Subsequent research, however, has demonstrated that listeners are sensitive to intracategory variation even with stimuli over which categorical perception can be easily obtained (e.g., McMurry, Tanenhaus, & Aslin, 2002; Miller & Volaitis, 1989; Pisoni & Tash, 1974).

In addition to the frequency and variability of the exemplars themselves, another distributional feature of the input to which learners are sensitive is the context in which exemplars occur. These contextual distributions can serve as cues to categorical distinctions. For example, Thiessen (2007) found that infants were more likely to use the categorical distinction between phonemes (such as /d/ and /t/) in word-learning contexts if they had previously seen the phonemes associated with different lexical forms (such as *diaper*, *doggy*, *tiger*, and *toothbrush*). This is an example of a phenomenon known as *acquired distinctiveness*: When two similar stimuli are paired with distinct contexts, learners distinguish between them more easily (Honey & Hall, 1989; James, 1890). This, in turn, makes it more likely that the learner will treat them as members of different categories, whether in terms of the learner's response to those stimuli (e.g., Edwards, Jagielo, Zentall, & Hogan, 1982) or in terms of detecting a change from one stimulus to the other (e.g., Thiessen & Yee, 2010).

Cue-Based Statistics

A final ability that falls under the umbrella of statistical learning is cue discovery and weighting. This occurs when learners detect that some perceptual feature of the input indicates the presence of a property that is not directly accessible from the perceptual characteristics of the input. For example, unlike the white spaces between words in print, word boundaries in speech are not marked by a consistent perceptual feature. Similarly, many internal characteristics, such as animacy or emotional state, cannot be directly observed. Instead, the presence of these properties must be inferred from the directly perceptible cues with which the properties are associated (e.g., Rakison & Lupyan, 2008). *Cue-based statistical learning* refers to the process through which infants discover which perceptible attributes of the input are correlated with attributes that are not directly perceptible, and how they learn to weight some cues more heavily than others (e.g., E. K. Johnson & Seidl, 2008).

In the context of statistical learning, the most widely studied example of cue-based learning is the discovery of acoustic cues to word boundaries such as pauses, phonotactics, and lexical stress (e.g., E. K. Johnson & Jusczyk, 2001). For example, words in English typically begin with a stressed syllable (Cutler & Carter, 1987), and adults treat stressed syllables as word onsets (e.g., Cutler & Norris, 1988). By the time they are 8–9 months of age, English-learning infants also use stress as a cue to word onsets (e.g., E. K. Johnson & Jusczyk, 2001). Strikingly, although younger infants fail to use stress as a cue (Thiessen & Saffran, 2003), they can be taught to do so through exposure to isolated words instantiating a regular stress pattern (Thiessen & Saffran, 2007). Similarly, infants can learn to use phonotactic regularities to segment fluent speech (Saffran & Thiessen, 2003). Critically, learners are able to identify these cue regularities when the cue functions probabilistically rather than deterministically (e.g., Gratton, Coles, & Donchin, 1992; Thiessen & Saffran, 2007).

This discussion of cues to word boundaries illustrates an important point about cue-based statistical learning: Learners generalize their knowledge about cues to novel settings. Once an infant has discovered that stress predicts word onsets, that knowledge will be applied widely. Infants will even apply their knowledge to settings in which it is incorrect, as in settings in which the correlation between cues and word boundaries is violated because a word is stressed on the second syllable (e.g., P. W. Jusczyk, Houston, & Newsome, 1999; Thiessen & Saffran, 2007). Of course, generalization can and does occur in conditional and distributional statistical learning (e.g., Maye, Weiss, & Aslin, 2008; Thiessen, 2011a). The generalization in cue-based statistical learning is especially notable, though, in that it alters subsequent

learning. This process has been referred to as "learning how to learn" (e.g., Harlow, 1949; Thiessen & Saffran, 2003; Yerkes, 1943). Once the perceptual features that cue underlying structure have been learned, these perceptual features change the way that learners detect structure in the future (e.g., Curtin, Mintz, & Christiansen, 2005).

Like other forms of statistical learning, cue-based learning is not limited to linguistic input. Cue-based statistical learning plays a role in visual tasks from simple flanker-style response tasks (e.g., Lehle & Hubner, 2008), to judgments about whether a particular pattern of motion corresponds to animate or inanimate objects (e.g., Rakison, 2005), to nonlinguistic auditory tasks (e.g., Holt & Lotto, 2006), and many other domains. In all of these domains, learners weight the use of a cue at least partially as a function of the strength of the probabilistic relationship between the cue and the critical category or response (e.g., Gratton et al., 1992; Thiessen & Saffran, 2004). In each of these domains, however, the critical feature of cue-based statistical learning is that the cue, once identified, serves to shape subsequent learning.

Many theories of cue weighting rely on attention to explain the increasing sensitivity to valid cues (e.g., Griffiths & Mitchell, 2008; Lehle & Hubner, 2008; Samuelson & Smith, 2000; Smith, Jones, & Landau, 1996). That is, once a cue has been identified, attention to the cue affects subsequent learning and performance. Attentional theories of cue-based learning are quite consistent with the fact that the salience of the cue appears to play an important role in how easily a cue is learned (e.g., Trabasso & Bower, 1968). Indeed, salience can be a more powerful factor than reliability in predicting learners' generalizations. In some instances in which a very subtle cue is reliable, learners will generalize on the basis of a less reliable but more noticeable cue (e.g., Emberson, Liu, & Zevin, 2009; Lidz et al., 2003). The emphasis on salience over reliability is heightened when learners are under time pressure (e.g., Lamberts, 1995). This should not be taken to mean that salience is necessarily separate from or in opposition to statistical learning. What a learner considers to be a salient property of the input is due in part (though certainly not entirely) to their prior experience with statistical structure of the environment (e.g., Honey & Hall, 1989). As this discussion indicates, a complete model of statistical learning should incorporate effects of, and influences on, attention. This is a point to which we return in the Linking Extraction and Integration Through Attention section.

Summary

Although it is clear that learners are sensitive to many different kinds of statistical relations in the input across a variety of input domains, it is unclear whether all three forms of statistical learning share the same underlying mechanism, as might be implied by the fact that all are referred to by the term *statistical learning*. Our goal in this review was to advance a memory-based framework for statistical learning that will clarify this mechanistic question. This framework is based on the combination of two processes that we term *extraction* and *integration*. Extraction is the process of identifying statistically coherent clusters (defined by conditional relations) of perceptual features and storing them in memory as discrete representations (such as word forms). Integration is the process of comparing across those clusters to identify commonalities and the central tendency of the input. As we argue, neither process in isolation is capable of accounting for the range of statistical learning phenomena described above. In combination, however, they enable a unified framework to explain the entire range of statistical learning.

Extraction and Conditional Statistics

Models of conditional statistical learning can be classified into two groups: boundary-finding models and clustering models. Boundary-finding models search for regions in the input where conditional relations between adjacent elements are (relatively) low and impute a boundary between units there. For example, some serial recurrent networks are trained to predict individual elements in a sequence of speech sounds on the basis of previous sounds. Word boundaries can be inferred at any region where predictability of the next sound falls below a threshold (e.g., Elman, 1990). Boundary-finding models do not represent or store the units (such as words) whose boundaries they discover; rather, they learn and represent the statistical relations between elements in the input. Clustering models, by contrast, store clusters of statistically related elements (e.g., Giroux & Rey, 2009). In word segmentation tasks, for example, clustering models store clusters of speech sounds with strong conditional relations in a lexicon of candidate word forms (see Orbán, Fiser, Aslin, & Lengyel, 2008, for an example of a clustering model in a nonlinguistic domain). Several different types of clustering models have been proposed in the statistical learning literature. Two of the most prominent approaches are chunking (Perruchet & Vinter, 1998) and Bayesian hypothesis testing (e.g., Frank et al., 2010).

Different clustering models invoke very different processes, and it is not yet clear which type of clustering model is the most faithful simulation of the human learning process (Frank et al., 2010). Assessing which formulation of the clustering approach is the best fit to human learning is beyond the scope of this review. However, all clustering models concur that the output of learning is a set of statistically coherent clusters that have been extracted from the input and stored in memory as discrete representations. In this section, we review the evidence that statistical learning does, in fact, result in the formation of this type of discrete representation. This evidence provides strong support for the claim that conditional statistical learning involves some form of extraction of coherent clusters (such as word forms) from the input. But although a process of extraction allows for sensitivity to conditional statistical relations, we argue that clustering models are poorly suited to explain sensitivity to distributional and cue-based statistical regularities. That is to say, a complete account of statistical learning requires some process in addition to extraction.

Evidence in Favor of Extraction

The primary claim of clustering models is that learning results in discrete representations that have been extracted from the input (e.g., words from a sentence, or shapes from a visual array). This claim has been examined directly on a number of occasions with linguistic stimuli in statistical learning tasks and is supported by evidence that humans treat the items they segment as lexical items. For example, infants accept words from the synthesized speech in English utterances after exposure to a stream of synthesized speech (Saffran, 2001). Similarly, infants and adults learn labels for novel

objects more easily when provided the opportunity to segment the labels from fluent speech (Graf Estes, Evans, Alibali, & Saffran, 2007; Mirman, Magnuson, Graf Estes, & Dixon, 2008). These results indicate that the objects that infants and adults identify via statistical learning are represented in the unitized manner consistent with word forms, as should be the case if clustering models are correct.

In clustering approaches, speech segmentation consists of synthesizing a set of elements (e.g., syllables) into larger units (e.g., words). This means that learners should differ in their response to subcomponents of a unit (e.g., *eleph* from the word *elephant*) as they become more familiar with the overall unit. As the learner becomes more familiar with the larger unit, the subcomponents embedded within that unit become less plausible candidate items to segment from the input. To put this in terms of linguistic materials, as a learner becomes more certain that *elephant* is a word, *eleph* becomes a less compelling candidate word; the word interferes with other potential word forms embedded within it. As the longer word form accrues more evidence, the embedded components become less plausible and are removed from the lexicon (Frank et al., 2010; Perruchet & Vinter, 1998). Although different clustering models instantiate this competition between clustered units and their embedded components differently, they share the prediction that these items compete to be extracted (e.g., Giroux & Rey, 2009; Orbán et al., 2008).

This competition between extracted units and embedded components is not present in boundary-finding models (e.g., Christiansen et al., 1998). Boundary-finding models identify boundaries on the basis of the probabilistic structure of the input. For example, with linguistic stimuli such as a sequence of syllables, these models learn the likelihood that two syllables will co-occur, and posit word boundaries at regions in the sequence where predictability is low. Increased exposure strengthens the model's knowledge of the probabilistic relations between all of the syllables within a word equally. This is a key difference between clustering and boundary-finding approaches that leads to an empirically testable prediction (e.g., Giroux & Rey, 2009). From a clustering perspective, as learners become more familiar with units (e.g., a word), they should become less able to distinguish subcomponents from within that unit (e.g., *eleph* from *elephant*) from a random configuration of elements; their knowledge of embedded subcomponents weakens due to interference as their familiarity with the overall unit increases. From a boundary-finding perspective, as the learner becomes more familiar with a unit, they should also become better at distinguishing embedded components from random configurations of elements. This is because boundary-finding models are storing the likelihood of transitions between elements rather than extracting words.

In laboratory experiments, humans conform to the predictions of a clustering account, rather than a boundary-finding approach: As exposure to the language increases, participants actually become less able to identify the sublexical embedded components within a word (Giroux & Rey, 2009). Indeed, this phenomenon is not limited to speech. Fiser and Aslin (2005) demonstrated the same process with visual input. In their experiments, participants were presented with a series of visual scenes composed of 12 shapes displayed simultaneously. These shapes were distributed into combinations that always occurred together, a kind of conditional statistic to which learners are sensitive (e.g., Fiser & Aslin, 2002). While participants learned the three-shape complexes, they failed to distinguish spurious pairs from pairs that were embedded in a complex. For example, if *star-triangle-square* and *moon-diamond-hexagon* were both true triplets, learners would fail to distinguish between *star-triangle* and *square-moon* (see Orbán et al., 2008, for further discussion). Fiser and Aslin (2005) refer to this as an "embeddedness constraint": As overarching structures are learned, knowledge of the constituent forms embedded in those structures is attenuated. The fact that this constraint can be seen for both sequential linguistic input and simultaneous visual input suggests that the extraction of units (as opposed to simply learning transitions between elements) is a domain-general feature of conditional statistical learning.

The Relation Between Extraction and Conditional Statistics

The extraction of statistically coherent clusters is informative with respect to the format representations in memory. It is important to note, however, that it is possible to define "statistically coherent" in a variety of ways. Several different models of conditional statistical learning (primarily with respect to word segmentation) have been suggested, many of which take advantage of different statistical metrics (see Frank et al., 2010, for an overview of several different clustering models). Most of these models fit extant human data reasonably well, so it is not yet possible to conclusively differentiate between them. All of these models, though, are meant to achieve sensitivity to *conditional* relations—that is, to identify clusters in the input whose elements strongly predict each other. This is important, because human learners are sensitive to conditional relations, and not simply identifying items in the input that frequently occur (e.g., Aslin et al., 1998). To illustrate how clustering models achieve sensitivity to conditional relations, we discuss two types of models: a chunking model and a Bayesian model. Although this is not an exhaustive review, the principles that enable these models to identify clusters that are coherent (not merely frequent) will provide an introduction to the logic that underlies clustering models more generally.

For many kinds of input, statistical coherence and frequency are confounded. As an example, consider the difference between words (syllable groupings that are coherent) and spurious syllable groupings formed across word boundaries (such as the grouping *tyba* in *pretty baby*). Because by definition the words *pretty* and *baby* will occur more frequently than the conjunction of these two words, the statistically coherent items are also the more frequent items. As such, before claiming that a model is sensitive to conditional probability, it is important to demonstrate that they are not simply responding to the frequency with which an item occurs. A stringent test of a model's ability to benefit from conditional statistics in linguistic input is to present the model with a language in which one set of words occurs twice as often as a second set of words. In this case, although all the words have stronger conditional relations than part-words, the part-words formed between the frequent words occur just as often as the uncommon words. Both adults and infants are capable of learning a language of this type (i.e., of distinguishing between words and part-words), based solely on conditional statistical relations rather than frequency (Aslin et al., 1998). Both chunking models and Bayesian models

are capable of identifying words when presented with a language of this type, though they succeed in somewhat different ways.

To segment words, chunking models such as Parser (Perruchet & Vinter, 1998) rely on three processes: activation, decay, and interference. When initially exposed to a string of syllables, Parser randomly groups them into chunks. These chunks receive a set level of activation when they are first created; as the model proceeds through the input stream, the activation of the chunks stored in memory decreases over time (unless the chunks are reencountered) due to the effect of decay. Groupings of the input that are less likely to occur (e.g., syllables that co-occur across word boundaries) are less likely to be chunked. Even when they are chunked, they are less likely to be chunked again (because they occur relatively rarely), and thus more subject to decay. Over time, then, the chunks that are highly active in memory begin to reflect the statistical structure of the input, because those chunks that occur more often (i.e., words) receive more activation, and decay less, than infrequent chunks.

In addition to the frequency of chunks, Parser (Perruchet & Vinter, 1998) is sensitive to conditional structure. After exposure to a string of speech, the words in Parser's lexicon are more highly activated even than spurious syllable groupings across word boundaries that occur equally often. This is due to the effect of interference. If a component element (a syllable) within a chunk occurs in a different chunk, the prior chunk suffers interference. Consider the effect of interference on Parser's segmentation of a language with four words, *diti*, *bugo*, *dapu*, and *dobi*, where the first two words occur twice as often as the second two words (e.g., Aslin et al., 1998; Thiessen & Saffran, 2003). Any possible spurious syllable grouping that Parser chunks (e.g., *tibu*) is only one of several possible part-words that the model might chunk that contains syllables that overlap with each other (e.g., *ti* in *tibu* and *tida*). Whenever a chunk is created that overlaps with another chunk, the strength of these chunks will decrease due to interference. Thus, over a long period of exposure, part-words will be less active than words because words suffer from interference from fewer other potential chunks (Perruchet & Vinter, 1998).

Bayesian models of statistical word segmentation invoke quite different processes, but yield a similar insight about the way in which an extraction process favors words over spurious groupings of syllables across word boundaries (and thus achieves sensitivity to conditional relations). Bayesian models of word segmentation are hypothesis-testing models: They formulate a set of hypotheses about the potential segmentations of a string of text and then assess the likelihood of those hypotheses (e.g., Brent, 1999; Goldwater, Griffiths, & Johnson, 2009). For example, given a string of text like *pretty baby*, a Bayesian model might formulate segmentation hypotheses like *pre tty ba by*; *prettyba by*; *prettybaby*; *pretty baby*; and so on. To see how a Bayesian model assesses the likelihood of these different hypotheses, consider Goldwater et al.'s (2009) lexical model. The model estimates the likelihood of each candidate word form, based largely on the number of times the word form occurs in the text (the model's prior probabilities also play a role and bias it toward shorter words over longer words). Then the model estimates the likelihood of the hypothesis by multiplying the individual probability of the words posited by the hypothesis. This means that a segmentation hypothesis with fewer words (such as *pretty baby*) will tend to have an advantage over a segmentation hypothesis with more words (such as *pre tty ba by*), because there will be fewer terms to be multiplied.

The use of probability explains why Bayesian models favor extraction of words over extraction of frequent part-words. Consider, again, the artificial language with four words, *diti*, *bugo*, *dapu*, and *dobi*, where the first two words occur twice as often as the second two words (e.g., Aslin et al., 1998; Thiessen & Saffran, 2003). Although the part-words formed from the frequent words (*tibu* and *godi*) occur just as often as the infrequent words, they are less likely to be segmented. This is due to the fact that the part-words must occur less frequently than the words from which they are formed. Every time a part-word occurs, the word from which it was formed necessarily occurs as well, but the word can occur without the part-word (e.g., when *ditu* is followed by *dobi*, the part-word *tibu* does not occur). Because of this, segmentation hypotheses consisting of real words will tend to be more likely than segmentation hypotheses including part-words. The competition between words and part-words occurs not due to activation and interference (as in chunking models) but due to the fact that words will tend to be evaluated as higher probability than part-words, because they occur more frequently.

Bayesian and chunking models of word segmentation differ on many dimensions, but they are both sensitive to the key difference between words and part-words in these artificial languages: Words occur with greater predictability than spurious combinations of syllables across word boundaries. Because clustering models are searching for a limited set of items to extract (as opposed to, for example, storing the entire speech stream intact in memory), words and part-words compete. The greater predictability of words means that these items have an advantage in the competition to be extracted. Although different clustering models simulate this competition in different ways, they converge on the suggestion that the search for a limited set of items to extract gives rise to sensitivity to conditional statistical information.

Extraction Alone Is Insufficient to Explain Statistical Learning

Clustering models of conditional statistical learning are a good fit to the general characteristics of human sensitivity to conditional structure (e.g., Frank et al., 2010; Giroux & Rey, 2009; Orbán et al., 2008). But despite the fact that these models are able to extract the same items from the input as human learners, they are insufficient—in isolation—to explain other aspects of statistical learning. This is because models of extraction are concerned only with how learners identify a set of consistent clusters in the input, and store those clusters in memory. However, the extraction a set of clusters (such as words) does not explain either distributional or cue-based statistical learning, which both require the ability to generalize from prior experience. In the case of distributional learning, a model must be able to generalize to a novel exemplar on the basis of prior experience with a series of exemplars varying along some dimension (or set of dimensions). For cue-based statistical learning, a model must generalize prior experience with a cue (e.g., phonotactic patterns) to novel settings. A learner who is capable only of extraction will fail to detect the common features among extracted items, and be unable to benefit from them.

Indeed, this means that clustering models are unable to model some aspects of word segmentation. This can be seen, for example,

if a clustering model such as Parser (Perruchet & Vinter, 1998) is presented with a sequence of words all exhibiting a phonotactic regularity. As an example, consider what would occur if Parser were presented with a segmentation stream containing four words all following an *str*-initial pattern: *strafing*, *straggler*, *straiten*, and *stranded*. Because the syllables within these words have strong conditional relations, Parser would segment all four words if given enough exposure to the input. If a new fifth word were introduced into the segmentation stream, Parser would segment it quickly as well, having previously identified the other four (e.g., Perruchet, Tyler, Galland, & Peereman, 2004). However, Parser would segment that fifth word just as quickly if the word were consistent with the *str*-initial pattern (e.g., *stripling*) as if it were not. Parser has no way of detecting the phonotactic regularity and applying it to subsequent segmentation. A model that only extracts word forms from fluent speech has no mechanism for comparing across the items it has extracted. By contrast, human learners quickly learn to take advantage of those kinds of phonotactic regularities to facilitate subsequent segmentation (e.g., Saffran & Thiessen, 2003; Thiessen & Saffran, 2007). Extraction provides a lexicon of candidate words, but it does not describe how the commonalities across those words are detected or used.

Extraction alone is similarly ill equipped to account for distributional statistical learning, because extraction does not assess the similarity among extracted items. Similarity is critically important for discovering categories (e.g., Madole & Oakes, 1999). The effect of similarity among prior exemplars can most clearly be seen in the fact that humans are sensitive to the central tendency, or prototype, of a set of exemplars. Given exposure to a series of individual exemplars, memory preserves individual details of some, if not all, of those exemplars (e.g., Hintzman, 1976; Hintzman, Block, & Summers, 1973). In addition to recognizing prior exemplars as familiar, participants will also often verify novel exemplars as familiar if they are close to the average of the collective exemplars (e.g., Nosofsky & Zaki, 2002; Roediger & McDermott, 1995). However, discovery of central tendency, category structure, and abstract concepts is only possible if learning involves some process that benefits from the similarity of stored exemplars. Models that do nothing but extract conditional-related elements are able to store exemplars, but they do not take full advantage of the information that they have stored.

Summary

Clustering models have been proposed to simulate human learning in many kinds of sequential learning tasks in which identification of conditional relations plays a role. These models extract clusters of conditionally related elements from the input and store them in memory. Although the majority of these models have focused on word segmentation (e.g., Goldwater et al., 2009; Perruchet & Vinter, 1998), they have also been shown to provide a good fit to other tasks, such as sequence learning or segmentation of visual arrays into smaller chunks (e.g., Orbán et al., 2008). Despite the success of clustering models in identifying conditional statistical structure, the process of extracting chunks of conditional-related elements is insufficient, on its own, to account for the full range of statistical learning phenomena. Extraction of a set of items does not account for distributional or cue-based statistical learning. In the next section, we discuss a complementary set of models that fail to acquire conditional statistical structure but that provide excellent sensitivity to distributional information by integrating information across stored exemplars.

Integration and Distributional Statistics

For many kinds of stimuli, statistical structure is not a function of the sequence or conditionalized probability in which they occur, but rather their distribution along a continuum of similarity. This is the case, for example, in learning phonetic categories. Relatively unpopulated regions along a perceptual continuum are likely to indicate a boundary between phonetic categories. Regions of perceptual space near a category boundary (e.g., between /p/ and /b/ along a continuum of voice-onset time) are ambiguous: Exemplars there are equally close to either category, so it is difficult to determine their category membership (Liberman et al., 1957). Due to communicative pressures favoring clarity, speakers produce relatively few exemplars in these regions, and therefore regions with a sparse number of exemplars provide a cue to category boundaries (e.g., Maye et al., 2002). To identify the category structure in the input, the order of the exemplars does not matter, nor is there any necessary conditional relationship between one exemplar (or category) and the next. Instead, the most important statistical property is the distribution of exemplars across the continuum: their frequency and variability in relation to other exemplars (e.g., Clayards et al., 2008; Maye et al., 2002). Models that provide sensitivity to distributional statistics, therefore, must take advantage of different kinds of regularities in the input. In particular, these models are dependent on some encoding of the similarity structure of the input. Using similarity, these models are able to integrate across exemplars to identify the central tendency of the input and to generalize to novel instances, two abilities that we believe are necessary for a complete description of statistical learning.

Similarity and Generalization

Modeling the effect of similarity on human learning has been conducted in a number of different ways. According to *exemplar memory* accounts, all prior exemplars are stored in memory (e.g., Hintzman, 1984; Nosofsky & Zaki, 2002). Sensitivity to central tendency occurs by summation across prior exemplars. A retrieval cue activates all memory traces simultaneously, weighted by their similarity to the cue (e.g., Hintzman, 1986). An alternative approach, a *distributed system*, does not represent individual memory traces. Instead, each experience affects a distributed set of units. The memory trace of a particular experience is represented by the change in the strength of interconnections between those units. In such a system (e.g., McClelland & Rumelhart, 1985), traces are "superimposed" over each other because each trace influences the connections between units. The primary difference between these approaches is that exemplar memory models preserve each experience independently, whereas in distributed models the collective impact of multiple traces is preserved. In both approaches, however, sensitivity to the central tendency of collected experience arises without explicit representation. For example, both kinds of models allow for sensitivity to prototypes and abstract categories to emerge as a function of the aggregation of multiple prior memories (e.g., Hintzman, 1986).

For ease of exposition, we describe the characteristics of long-term memory models primarily in terms of exemplar memory models such as Hintzman's (1984) MINERVA 2. Framing the discussion in terms of memory trace models does not represent a necessary theoretical commitment; with relatively minor alterations, the framework we describe could be altered to incorporate a distributed long-term memory system, as in a connectionist architecture. However, we favor an exemplar memory model framework—at least for descriptive purposes—because it establishes an immediate point of connection with conditional statistical learning. The exemplars extracted via conditional statistical learning can feed into the discrete representations used in exemplar memory models. But our central theoretical claim does not relate to the superiority of exemplar memory models over distributed models of long-term memory. Instead, our central argument is that accounts of statistical learning are incomplete unless they incorporate two processes available in models of long-term memory: sensitivity to similarity and integration across multiple exemplars.

MINERVA 2 is a model intended to abstract the central tendency of a set of exemplars. It does so by comparing a current exemplar (whatever perceptual stimulus is being experienced) with exemplars previously stored in long-term memory, and returning a weighted average of the stored exemplars. The average is weighted by similarity: The prior exemplars that are most similar to the current exemplar contribute most strongly to the response generated from long-term memory. In MINERVA 2 (Hintzman, 1984), a memory trace is encoded as a vector of features with values ranging between −1 and 1; 1 can be thought of as the presence of the feature, −1 as its absence, and 0 as an indeterminate value. The similarity between two percepts can be computed as the sum of the product of cross-multiplication between the vectors, divided by the number of nonzero features. For example, 1, −1, 1, −1 is a four-feature percept that is maximally similar to 1, −1, 1, −1, yielding a similarity value of 1. It is maximally dissimilar to −1, 1, −1, 1, yielding a similarity value of −1. Note that in a distributed memory model, the same effect would be realized as overlap in the pattern of activation caused by the two experiences.

When a retrieval cue (or *probe*) is sent to previously stored traces in long-term memory, the weighted average of all the stored traces is returned. This average (the *echo*) has two characteristics: intensity and content. The intensity is related to the similarity between the probe and prior experiences; the more similar the probe is to previous experiences, the more active the echo is. The content of the echo is the sum (normalized by the number of traces) of the activity of the traces in memory, weighted by their activity. Traces more similar to the probe will contribute more to the echo than dissimilar traces. Depending on the specificity of the probe, many or few traces may be highly activated, and depending on the homogeneity among (especially highly) active traces, the content of the echo may be ambiguous, or consistent. The returned echo, then, has the property of embodying the weighted average of all of the activated traces, leading MINERVA 2 to be able to emulate prototype formation and schema abstraction (Hintzman, 1986).

Like many models of long-term memory, MINERVA 2 is characterized by sensitivity to similarity structure in the input (achieved via the vector coding scheme) and an ability to integrate information across prior exemplars (achieved via the probe/echo process). This enables MINERVA 2 to identify the central tendency of prior examples (Hintzman, 1984). Just as importantly, it allows the model to generalize on the basis of prior experience (Hintzman, 1986). This is because when the model is presented with a stimulus, the most similar prior experiences are activated. Even if the stimulus is novel, some prior experiences will be activated. If the activated prior exemplars have consistent features, those features will be returned from long-term memory, and guide the response to the (novel) stimulus to be similar to that of other, similar experiences in the past. This allows MINERVA 2 to benefit from commonalities across prior experiences, unlike those models of conditional statistical learning that merely extract items from the input but fail to detect their commonalities.

Similarity Sensitivity and Distributional Statistics

The processes embodied in MINERVA 2 (and many other models of long-term memory) yield sensitivity to distributional statistical information such as frequency and variability. For an example of how a similarity-based memory trace system gives rise to sensitivity to distributional information, consider learning to use phonemic contrasts in word-object association tasks. Thiessen (2007; see also Thiessen & Yee, 2010) found that 14-month-old infants' use of phonemic contrasts in a word-object association task is facilitated by exposure to those phonemes in distinct lexical forms. Infants often fail to use phonemic differences in such tasks; after habituation to a novel object labeled *daw*, infants accept *taw* as a label for the object (e.g., Stager & Werker, 1998). Thiessen (2007) found that exposure to the phonemes /d/ and /t/ in distinct lexical contexts lessened children's willingness to accept minimal pair labels interchangeably. After exposure to *dawbow* and *tawgoo*, children no longer accepted *taw* as a label for the object previously called *daw*. This was not simply due to increased exposure to the sounds /d/ and /t/, as hearing the sounds in *dawgoo* and *tawgoo* (an identical lexical context) did not facilitate children's performance. Instead, exposure to the sounds in distinctive contexts appears to be critical. This may be related to the phenomenon of acquired distinctiveness: Two similar percepts (in this case, /d/ and /t/) become more easily distinguishable if they are paired with different contexts or consequences (e.g., Hall, 1991).

MINERVA's vector coding can be applied in a straightforward manner to the stimuli (*dawbow*, *tawgoo*, *daw*, and *taw*) used in Thiessen's (2007) experiment (see Table 1). To do so, we use a 16-feature vector, with the first eight features coding the first syllable, and the next eight coding the second syllable. Within each syllable, the first four features describe the voicing status of the consonant, and then whether it is bilabial, alveolar, or glottal

Table 1
Vector Coding Scheme for Daw, Taw, Dawbow, Tawgoo, and Dawgoo

	Features															
Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Daw	1	0	0	1	1	−1	−1	−1	0	0	0	0	0	0	0	0
Taw	−1	0	0	1	1	−1	−1	−1	0	0	0	0	0	0	0	0
Dawbow	1	0	0	1	1	−1	−1	−1	1	1	0	0	−1	1	0	1
Tawgoo	−1	0	0	1	1	−1	−1	−1	1	0	1	0	−1	1	1	1
Dawgoo	1	0	0	1	1	−1	−1	−1	1	0	1	0	−1	1	1	1

(because this demonstration requires no dentals, fricatives, or liquids, these features are omitted). The next four features describe the vowel on the basis of whether it is front/back, rounded/unrounded, high/low, and long/short (in the front/back feature, for example, 1 indicates a front vowel and −1 indicates a back vowel). Though this coding system is loosely based on classical linguistic feature systems (e.g., Chomsky & Halle, 1968), it should be viewed solely as a notational convenience. It does not imply that learners actually represent speech in a featural, abstract manner (for discussion, see Thiessen, 2011a; Thiessen & Yee, 2010;). Instead, it is meant simply to capture in a quantitative way the fact that some lexical forms are more similar than others, information that is lost in many models of conditional statistical learning. The training regime from Thiessen (2007) can be simulated by two different sets of traces. The first set consists of *dawbow* and *tawgoo*, and the second set consists of *dawgoo* and *tawgoo*.

Consider what happens when these two sets of traces are probed with the test items infant participants heard: *daw* and *taw*. For the memory set consisting of *dawgoo* and *tawgoo*, the probe of *daw* activates traces of *dawgoo* more than traces of *tawgoo*; the probe of *taw* has precisely the opposite effect. However, because *dawgoo* and *tawgoo* differ only in the voicing of their first consonant, the echo that is returned in response to *daw* differs from that returned to *taw* only with respect to the single feature of the voicing status of the first consonant. The *daw* echo is weighted more toward voicing on this single feature; the *taw* echo is weighted more toward voicelessness. In all other respects, the *daw* echo and the *taw* echo are identical. In contrast, for the set of traces consisting of *dawbow* and *tawgoo*, the echoes returned to probes of *daw* and *taw* are more distinct. In this case, the echoes differ not only in the feature representing voicing for the consonant in the first feature but also in the features corresponding to place of articulation for the consonant in the second syllable. The *daw* echo contains traces of bilabial activation (due to its association with the second syllable *bow*), whereas the *taw* echo is more weighted toward glottal articulation. Thus, in the case in which phonemes have been experienced in different lexical contexts, probes containing those phonemes result in more differentiable echoes. Unlike chunking models, the MINERVA 2 model (Hintzman, 1984) captures the phenomenon that experiencing phonemes in different lexical contexts should promote differentiation of those phonemes (e.g., Thiessen, 2007; Thiessen & Pavlik, in press).

In addition to identifying categories from the distribution of exemplars across contexts, memory trace models are able to use the frequency of different exemplars to learn category boundaries. Consider Maye et al.'s (2002) demonstration that the distribution of exemplars along a continuum of voicing influences their discriminability. In a unimodal distribution in which most of the exemplars occur near the center, either endpoint on the continuum (voiced or voiceless) is equally near to the largest mass of exemplars. Therefore, the echo created in response to the probe of either endpoint is fairly similar, with an ambiguous central value for voicing. The result is quite different if the learner has been exposed to a bimodal distribution, with one mass of exemplars near the voiceless endpoint, another near the voiced endpoint, and comparatively few in the middle. In this case, the echo to the voiceless probe is primarily informed by the large mass of voiceless exemplars. The echo in response to the voiced probe is primarily influenced by the large mass of voiced exemplars. Therefore, the two echoes are more distinct than in the unimodal context, in which both echoes are influenced by the mass of central exemplars.

Note that the input to the learning process in the Maye et al. (2002) experiment is simplified from actual language in a number of ways. First, the exemplars varied along only a single feature (voicing). Second, all of the exemplars fell into discrete steps along this continuum, such that there were many identical exemplars. Third, the same speaker produced all of the exemplars. Finally, infants were not required to generalize their knowledge to new contrasts (though see Maye et al., 2008). As such, it is not clear whether the distributional learning made possible by MINERVA 2's sensitivity to central tendency (Hintzman, 1986) would suffice for learning from naturalistic language input. As such, the claim here is not that trace memory is sufficient for language learning (though see P. W. Jusczyk, 1993). Rather, the claim is a more limited one: that sensitivity to similarity, and integration across prior experience, is sufficient to describe the kinds of distributional statistical learning phenomena seen in laboratory experiments with infants, adults, and animals. Although these processes are also undoubtedly necessary for natural language acquisition, they are unlikely to be sufficient—models of the acquisition of natural language are likely to require additional processes.

Finally, the process of integration across exemplars instantiated in models like MINERVA 2 (Hintzman, 1984) also yields sensitivity to variability. The discovery of nonadjacent relations provides an example of this. Discovering nonadjacent relations—such as discovering that *ko* predicts *be* in novel words like *kotibe*, *kosube*, *komabe*, and *kolabe*—is more difficult than discovering adjacent relations (e.g., Creel et al., 2004). One factor that facilitates the discovery is the variability of the element intervening between the nonadjacent relationship. When exposed to a string of A-X-C items, where A predicts C, both infants and adults are better able to detect the nonadjacent A–C relationship when the intervening X item is more variable (Gómez, 2002). When there are very few syllables that can fill the X position, learners fail to detect the nonadjacent relationship. The MINERVA 2 framework provides a straightforward mechanistic account to explain this effect of variability (for a more extensive discussion, see Thiessen & Pavlik, in press).

Consider what would occur if long-term memory contained four exemplars: *kotibe*, *kosube*, *komabe*, and *kolabe*. A subsequent probe to memory that is similar to these items (such as the first syllable *ko*) would activate all of these traces. The resultant echo returned from memory would be strongly consistent about the information that is constant across all of these traces. The inconsistent information would tend toward canceling itself out (e.g., Hintzman, 1986). Therefore, upon a probe of *ko*, this memory system would return *ko*X*be*, where the X represents information that is inconsistent across the traces and thus not strongly endorsed in the echo. Furthermore, inconsistent elements are more likely to cancel out as there are more exemplars with different elements. If a learner experiences only two strings (e.g., *komabe* and *kolabe*), both of the X elements contribute fairly strongly to the echo, especially for those features for which they are consistent. If a learner experiences many unique strings, none of those individual exemplars contributes as strongly to the overall echo. Only the information that is consistent across all of the traces is strongly

represented. In highly variable input, the information about the X element is likely to be lost or cancelled out, and would not contribute to the judgment of similarity between prior exemplars and novel strings obeying the A-C regularity.

On the surface, tasks such as detecting nonadjacent relations, learning category boundaries, and using contextual information to disambiguate phonemes share little in common. One advantage of modeling is that it provides an opportunity to assess whether the same process can potentially explain behavior in very different tasks. Although the MINERVA 2 architecture—and more generally, models of information integration—are not new (e.g., Hintzman, 1984; McClelland & Rumelhart, 1985), their application to statistical learning phenomena has been relatively limited (though see Adriaans & Kager, 2010). A novel claim of this framework is that integration across exemplars can explain all of the distributional statistical learning phenomena discussed above. If this claim is correct, it should be possible to model all of these phenomena using the same computational model (for a demonstration using an updated version of the MINERVA 2 architecture, see Thiessen & Pavlik, in press).

Extraction and Integration: Two Incomplete Approaches

Extraction and integration each account for an aspect of statistical learning. Extraction is the process via which statistically coherent clusters of elements are stored in memory. Integration allows learners to discover the central tendency of the exemplars stored in memory, and benefit from their similarity and distribution. One possible view of these processes is that they operate independently and in isolation, accounting for separate aspects of statistical learning. As the discussion in the previous section indicates, models of long-term memory are able to simulate distributional learning by integrating information across multiple exemplars. However, these models are typically not designed to simulate learning from conditional relations. Memory trace models such as MINERVA 2 (Hintzman, 1984), or the McClelland and Rumelhart (1985) model, do not have a principled way to segment sequential input into discrete chunks, and they often assume that the input for learning comes presegmented (e.g., P. W. Jusczyk, 1993). As such, these models are agnostic with respect to the processes that lead to extraction of coherent chunks from continuous input.

Similarly, many models of extraction are limited in their ability to benefit from similarity structure. For example, chunking models of statistical learning such as Parser (Perruchet & Vinter, 1998) have no way of detecting consistent phonological cues that occur in the words that they segment. If Parser were presented with a sequence of verbs all ending in *–ed*, it may, depending on the length of exposure and statistical structure of the input, segment all of the verbs successfully. However, Parser has no way of deducing that the novel verb is also likely to end with the same *–ed* suffix upon presentation of a novel verb stem, because Parser has no mechanism for comparing across chunks and identifying these similarities. For Parser to take advantage of these kinds of word-form regularities, they must be coded in advance (e.g., Perruchet & Tillmann, 2010). This difficulty is not unique to a chunking approach. Models of segmentation via conditional statistical relations often require that other cues (such as phonological cues) be built in ahead of time, rather than learned over the course of exposure to the input (e.g., Christiansen et al., 1998). These models assume (explicitly or implicitly) that the processes underlying the extraction of word forms from fluent speech can be modeled independently of the process of integration of information across those word forms.

As this discussion indicates, most models of statistical learning choose to simulate conditional and distributional statistical learning separately (though see Adriaans & Kager, 2010, for an exception). This is consistent with the possibility that conditional and distributional statistical learning operate separately. Such a proposal accounts for conditional statistical learning (accomplished via extraction) and distributional statistical learning (accomplished via sensitivity to similarity across memory traces). Each kind of process performs a separate task, and can be modeled quite successfully in isolation (e.g., Giroux & Rey, 2009; Hintzman, 1986; McClelland & Rumelhart, 1985; Perruchet & Vinter, 1998). However, if these processes are entirely separate, then there is no opportunity to account for cue-based statistical learning. Models of extraction have no way of discovering the phonemic regularities that come to play an important role in word segmentation (e.g., E. K. Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003). Similarly, models of integration lack a principled approach to apply the knowledge they acquire to word segmentation (e.g., P. W. Jusczyk, 1993). These difficulties suggest that an approach in which extraction and integration are conceptualized completely independently is inadequate. Instead, our perspective is that these two systems are deeply intertwined.

We believe that it is possible to create a framework that accounts for the full range of statistical learning phenomena only by combining extraction and integration in a single approach. From our perspective, the processes of integration and extraction bidirectionally influence each other. One route of influence is that the exemplars segmented from continuous input (via the process of extraction) serve as the input for the process of integration. This is consistent with the assumptions made by many models of long-term memory, which operate over exemplars that have previously been segmented from the input but which are agnostic with respect to the process via which the input is segmented (e.g., Hintzman, 1984; P. W. Jusczyk, 1993). In the domain of word segmentation, for example, this perspective suggests that conditional statistical cues allow infants to extract a set of candidate word forms, even before they have discovered phonological cues to word boundaries (Thiessen & Saffran, 2003). Once these word forms are stored, it is possible to integrate information across them and discover the phonological regularities that are consistent across these word forms, such as, in English, word-initial lexical stress (Thiessen & Saffran, 2007).

The second route of influence between the processes of extraction and integration is that the regularities discovered via the process of integration serve to inform subsequent extraction. Consider word segmentation again: It is a well-established finding that discovering phonological regularities changes subsequent segmentation (e.g., Saffran & Thiessen, 2003; Thiessen & Saffran, 2007). Indeed, many phonological cues, once learned, are weighted more heavily than the conditional relation between syllables (E. K. Johnson & Jusczyk, 2001). For example, once English-learning infants have discovered that stress is correlated with word-initial position, they will segment sequences like *TARis* (from *guiTAR is*)

from fluent speech, despite the strong conditional relation between the syllables in the word *guitar* (P. W. Jusczyk et al., 1999). The fact that these phonological cues exert a greater influence on segmentation than conditional statistics has often been interpreted to mean that segmentation via phonological cues is driven by a different process than sensitivity to conditional relations (e.g., E. K. Johnson & Seidl, 2008).

The suggestion that use of phonological cues arises from a different process than sensitivity to conditional relations is a natural extension of the view that segmentation via statistical cues requires the calculation of transitional probabilities. Transitional probabilities, by definition, are sensitive only to the conditional relations between syllables, and not to any kind of phonological regularities (e.g., E. K. Johnson & Jusczyk, 2001; Saffran et al., 1996). However, as we demonstrate in the next section, sensitivity to phonological cues can arise from the same segmentation system as sensitivity to conditional relations. This is only possible so long as the processes of extraction and integration are synthesized into a complete framework, which underscores the importance of considering these processes in combination rather than in isolation.

Summary

There are a variety of statistical regularities in which the critical statistical structure is a *distributional* structure: It relates to the frequency or variability of exemplars in the input. Models of conditional statistical learning—especially word segmentation—do not typically take advantage of this kind of distributional information (e.g., Christiansen et al., 1998; Frank et al., 2010; Perruchet & Vinter, 1998; though see Adriaans & Kager, 2010). Doing so requires a learning mechanism that is capable of taking advantage of the similarity between exemplars and identifying their central tendency. Models of long-term memory are ideally suited to this purpose because they have the ability to integrate information across exemplars, accentuating consistent information while de-emphasizing conflicting aspects of the exemplars. This process of integration allows these models to simulate a wide variety of distributional statistical learning phenomena, including discovering category boundaries, learning to use phonemic contrasts for word learning, and identifying nonadjacent relations (for more extensive discussion, see Thiessen & Pavlik, in press).

Despite the success of models of long-term memory in simulating distributional statistical learning by integrating across exemplars, these models typically ignore the process via which exemplars are extracted from the input (e.g., Hintzman, 1984; P. W. Jusczyk, 1993). The fact that most models and theoretical proposals consider extraction and integration separately underestimates the potential of statistical learning. Models of distributional learning are incomplete because they fail to provide a principled account of the origin of the exemplars across which they integrate information, and models of conditional learning are incomplete because the regularities that are discovered via distributional learning (e.g., phonological regularities) are never used to constrain subsequent extraction. Only a unified framework—one that combines the processes of extraction and integration—can remedy this.

Combining Integration and Extraction for Word Segmentation

Statistical learning has been researched most thoroughly with respect to word segmentation. Because of this, word segmentation provides the best domain to assess any potential account of statistical learning. Our goal was to assess the potential for an account of statistical learning that combines the processes of extraction and integration, so we discuss this account in the context of word segmentation. We propose that a framework that incorporates both of these processes will be able to account for a wider variety of results than any prior account of statistical learning. Prior models and theoretical accounts of statistical learning have been single-process accounts: They are focused on either extraction or integration. Modeling these processes separately, and exploring them in laboratory tasks, has resulted in important advances in our understanding of statistical learning. But studying these processes separately fails to take into account how they influence each other, and how this interaction changes across development.

Indeed, even in the same task, infants show different weightings of information available via extraction and integration as a function of age. When presented with a language for which conditional statistical information and phonological cues (discovered via the process of integration) conflict, younger infants favor conditional cues and older infants favor phonological cues (e.g., Thiessen & Saffran, 2003). Although prior laboratory experiments have been tremendously informative in mapping infants' weighting of different sources of information across age (e.g., E. K. Johnson & Jusczyk, 2001; P. W. Jusczyk et al., 1999), these experiments do not provide a general account of when infants will favor information that is derived from a single stimulus (i.e., conditional statistics), and when they will favor information that is derived from integration information across multiple exemplars (such as phonological regularities). This highlights the necessity of an account that incorporates both the processes of extraction and integration.

As discussed previously, there are multiple possible computational implementations of both extraction and integration. To implement the process of integration in this framework, we continue to use the MINERVA (Hintzman, 1984) exemplar memory approach described above. For an implementation of the process of extraction, we have chosen to use Parser, a chunking model (Perruchet & Vinter, 1998). Although these implementation choices are not arbitrary, they should not be taken as a claim that these are the only possible implementations for an account of statistical learning that combines extraction and integration. Instead, these choices were made because of the straightforward manner in which it is possible to combine a chunking model with an exemplar memory model: The chunks extracted via the process of chunking are fed to long-term memory as the exemplars to be integrated across. The novel contribution of this framework is not an examination of a particular computational instantiation of the processes of extraction and integration, but rather an exploration of how these processes work in concert.

Benefits of a Chunking Model of Extraction

There are three advantages to a chunking model (as an implementation of the process of extraction) for the present framework: Chunking provides an adequate fit to human performance in con-

ditional statistical learning tasks; it does so without calculating transitional probabilities; it can be combined easily with exemplar memory models (our choice as an implementation of the process of integration). A brief discussion of Parser (Perruchet & Vinter, 1998) will help to illustrate these points. Parser segments words due to the effects of activation, decay, and interference. When exposed to a sequence of syllables, Parser randomly groups them into chunks. This grouping is thought to reflect the action of attention; only those syllables that are simultaneously held in attention are chunked. Over time, the activation of these chunks decays unless the chunks are subsequently encountered again, in which case their activation is increased. If a syllable within a chunk occurs in a different chunk, the prior chunk suffers interference and loses activation. On average, words will be encountered more than spurious groupings across word boundaries (such as *tyba* from *pretty baby*), so the effect of interference will be comparatively larger for spurious groupings than for real words. As the model receives more exposure to the language, the chunks that are most active come to reflect the statistical structure of the input, because the chunks that occur more often (i.e., words) receive more activation, and decay less, than spurious groupings.

Chunking models provide a good fit to human data from word segmentation tasks. Perhaps the most compelling point of convergence between the chunking approach and human data is that chunking models are constructing a set of potential word forms (i.e., chunks). As discussed previously, this also appears to be what human learners are doing during word segmentation tasks. For example, segmentation appears to yield representations of the segmented items as a single unit, as opposed to a set of associations of the elements (e.g., syllables) within the unit (Fiser & Aslin, 2005; Giroux & Rey, 2009). Moreover, the items that are learned in a segmentation task appear to be lexicalized, in that they are potential labels for novel objects (Graf Estes et al., 2007; Mirman et al., 2008). The fit between chunking models and human performance suggests that Parser is a good starting point as a computational instantiation of the process of extraction.

For the purposes of combining extraction with integration to provide an account of cue-based statistical learning, Parser (Perruchet & Vinter, 1998) has a second advantage: It segments without calculating transitional probabilities. Transitional probabilities describe the conditional relation between syllables (e.g., Aslin et al., 1998; Saffran et al., 1996). Transitional probability models segment on the basis of these conditional relations, for example, by inserting word boundaries at regions where the probabilities fall below a certain threshold (e.g., Frank et al., 2010). Because transitional probabilities are, by definition, only influenced by the conditional relation between syllables, segmentation via transitional probabilities is necessarily insensitive to phonological regularities such as phonotactics, coarticulation, or lexical stress (E. K. Johnson & Jusczyk, 2001). A model that segments via transitional probabilities does not know, for example, that certain phoneme pairs are much more likely to occur at word boundaries than within a word; instead, it segments solely on the basis of the likelihood that syllables co-occur. Because of this, theories of segmentation that incorporate both phonological regularities and transitional probabilities often do so using stages of processing. First the speech stream is segmented by phonological regularities, and then transitional probabilities are used to segment candidate words from ambiguous regions of the input where phonological regularities are uninformative (e.g., Mersad & Nazzi, 2011; Shukla, Nespor, & Mehler, 2007).

By contrast, according to chunking models, learners are not calculating transitional probabilities in order to segment speech. Instead, the items that are extracted from speech are stored as a function of activation, interference, and decay. Thus, from the perspective of chunking accounts, statistical measures like transitional probability are understood as a description of the statistical structure of the input, but play no part in the process of learning. To illustrate this point, consider backward and forward transitional probabilities. Humans are sensitive to conditional relations in both a forward-going (*Batman* predicts *Robin*) and a backward-going (*the* does not strongly predict *cat* going forward, but *cat* has a strong "backward prediction" to *the*) direction (e.g., Onnis & Thiessen, in press; Peluchi, Hay, & Saffran, 2009). This is consistent with the principles of a chunking model such as Parser. As an example, if Parser has segmented an artificial language in such a way that *bapi* is a candidate lexical item, this chunk suffers as much interference from an item like *balu* (which lowers the forward transitional probabilities of *bapi*) as it does from an item like *gopi* (which lowers the backward transitional probabilities). From a chunking perspective, sensitivity to transitional probability in both directions arises naturally from interference (Perruchet & Desaulty, 2008). Rather than attempting to calculate both the forward-going and backward-going transitional probabilities (and choosing a method to combine these different metrics), Parser (Perruchet & Vinter, 1998) achieves segmentation by dint of processes that do not require calculation of transitional probabilities. This means that it may be possible for chunking models—unlike transitional probability models—to achieve sensitivity to phonological cues with the same process that yields sensitivity to conditional relations (Perruchet & Tillmann, 2010).

Indeed, the fact that it is possible to achieve segmentation without explicit computation of transitional probabilities has important implications. Although the experimental stimuli in statistical learning experiments are often described in terms of transitional probabilities, it is not clear what (if any) statistic is responsible for learning in these experiments. Saffran et al. (1996) were careful to note that transitional probability is only one of many possible descriptive statistics, and does not necessarily map onto the underlying computations of learners. Subsequent researchers have attempted to determine what kinds of explicit computations best capture learners' sensitivity to statistical structure (e.g., Aslin et al., 1998; Frank et al., 2010). For example, some authors have suggested that *mutual information* may better capture learners' statistical intuitions than transitional probabilities, because it captures the strength of the relationship between X and Y in both forward-going and backward-going directions (e.g., Brent, 1999; Redington, Chater, & Finch, 1998). Chunking models sidestep this debate about the statistical computations underlying learning, because they are not computing explicit statistics (Perruchet & Vinter, 1998). Rather, from this perspective, sensitivity to conditional statistical information arises from memory-based processes such as activation, interference, and decay.

A final advantage of a chunking model such as Parser (Perruchet & Vinter, 1998), with its emphasis on processes involved in memory, is that it can be combined with a model of long-term memory in a straightforward manner. One link between them is that the chunks that arise from the process of extraction can be fed

to long-term memory. There, the features that are consistent across those chunks can be identified through the process of integration. The knowledge that is accumulated in long-term memory can then influence subsequent chunking. This is possible because chunking does not depend on transitional probabilities to group and extract elements from the input. Rather, it relies on attention, which can be weighted toward different groupings as a function of prior experience (Perruchet et al., 2004). Parser's incorporation of attention as the mechanism that guides extraction means that it is flexible enough to benefit from the information stored in long-term memory. This is critically important; some method of benefiting from long-term memory is necessary for any successful combination of the processes of extraction and integration.

Linking Extraction and Integration Through Attention

In isolation, the process of extraction explains conditional statistical learning, and the process of integration explains distributional statistical learning. But neither process, alone, is able to account for cue-based statistical learning: the fact that prior experience enables learners to identify phonological regularities and use those regularities to constrain subsequent extraction (e.g., Lew-Williams & Saffran, 2012; Saffran & Thiessen, 2003; Thiessen & Saffran, 2007). By combining the processes of integration and extraction in a single framework, we believe that it is possible to explain cue-based statistical learning. This is possible because both memory trace models (our choice for a computational implementation of integration) and chunking models (our choice for a computational implementation of extraction) emphasize the role of attention in learning. Memory trace models typically incorporate differences in attentional weighting to explain increasing reliance on more reliable cues (e.g., Griffiths & Mitchell, 2008; Kruschke, 2001). In the word recognition and phonetic structure acquisition model (P. W. Jusczyk, 1993), for example, infants learn the features of their language from exposure to the distribution of sounds in their native language. On the basis of this distribution, the learner's attention is weighted toward features that occur across a wide variety of memory traces. That is, the echoes that result from probes of long-term memory traces guide the learner's attention. This leads infants to prefer the sound pattern of their native language to the sound patterns of foreign languages (P. W. Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993).

Chunking is similarly dependent on attention: Elements in the input are only chunked together when they are simultaneously held in attention. This raises the possibility that attention provides an avenue via which the information stored in memory traces can influence subsequent chunking behavior. Chunking models necessarily incorporate attention, given the proposed centrality of attention to the ability to chunk. However, attention in these models is often not explicitly simulated or does not behave in a systematic manner. Parser is no exception in this regard. Attention in Parser is deployed randomly, grouping together between one and three perceptual primitives. To interface with a set of memory traces, and to provide a better model of human learning, attention must influence chunking in a more orderly fashion. This would be possible if the distribution of attention (and thus, chunking) were influenced by prior experience. For example, most content words in English show word-initial stress (e.g., Cutler & Carter, 1987). English speakers, and English-learning infants, eventually become sensitive to this regularity and use it as a cue to subsequent word segmentation (e.g., E. K. Johnson & Jusczyk, 2001; P. W. Jusczyk, Cutler, & Redanz, 1993). Presently, Parser has no way of discovering and taking advantage of this kind of information. But this would be possible if Parser allowed similarity across prior chunks to influence attention in subsequent chunking.

If attention behaved systematically in Parser, then the size of chunks would not be entirely determined by chance. Instead, Parser could develop a bias to deploy attention to maximize the similarity between current chunks and prior exemplars. For example, if Parser were learning English, the prototypical word would be stress-initial. This is due to the fact that if Parser had successfully identified several words, any probe to long-term memory would return an echo that conveys clear information only on those features that were consistent across the majority of the (highly active) traces. If Parser knew the words *BAby*, *DOGgy*, *TAble*, and *SHOE*, the phonemic information would not be consistent across traces, but the stress pattern would be consistent. A probe to a lexicon like this would return an echo that is agnostic with response to phonemic identity but strongly consistent with the pattern that word-initial syllables are stressed. This could serve as a signal to bias attention to begin a chunk on a stressed syllable. Note that this approach provides a unified mechanism to explain infants' use of multiple different cues. Segmentation via conditional statistical information and via lexical stress arises from intimately related processes: chunking and the similarity across chunks. This is in contrast to conceptualizations in which phonological cues and "statistical" cues have been envisioned as competitors, arising from different processes (e.g., E. K. Johnson & Jusczyk, 2001; E. K. Johnson & Seidl, 2008).

This proposal about the effect of prior experience on attention is qualitatively different from the effect of experience in the original instantiation of Parser (Perruchet & Vinter, 1998). In Parser's original architecture, experience served to make chunks longer, as larger and larger segments of the input can be chunked as perceptual primitives. In our conception, prior experience helps to constrain the nature of chunks and makes them more similar to those chunks that have been previously experienced. Indeed, this warping of chunks toward prior experience is necessary to account for the fact that segmentation is influenced by the word-likeness of potential word forms in a speech stream. Boundary-finding models account for this by allowing several cues, such as stress and phonotactics, to influence the location of word boundaries (e.g., Christiansen et al., 1998). Parser can account for this phenomenon if chunks are constrained to be wordlike (i.e., to respect acoustic regularities) as demonstrated by Perruchet and Tillmann (2010). Perruchet and Tillmann altered Parser such that chunks were constrained to match estimates of wordlikeness (such that wordlike segments were more likely to be initially chunked). This proposal is similar to our framework, with one important exception: In Perruchet and Tillman's extension of Parser, ratings of wordlikeness are entered directly into the chunking algorithm. In our framework, learning about wordlikeness occurs from the experience of chunking itself. In that way, the learner's prior experience influences what they are likely to chunk in the future.

The effect of prior experience on attention is applicable to a wide variety of cues, including lexical stress and phonotactics. If one of the sequences in the speech stream matches a previously

experienced chunk, then that memory is activated most strongly and guides segmentation (e.g., Bortfeld, Morgan, Golinkoff, & Rathbun, 2005). If the speech stream contains novel words, no single trace is predominantly activated. Novel incoming speech activates many or all of the traces in memory, graded by the similarity of the input to prior chunks (e.g., Goldinger, 1998). Elements of those words that conflict cancel each other out, and only the information that is consistent across words is returned from long-term memory (e.g., Hintzman, 1984). In the case of stress, this highlights a word (or chunk) initial cue. But the interaction between traces and chunking can just as easily identify cues to word-final position. Consider an infant who has previously chunked the lexical forms *typing*, *singing*, *laughing*, and *kissing*. Upon being presented with the novel string *jumping jack*, the infant should avoid the segmentation *jump ingjack.* The infant's prior experience with *ing* is activated by the presentation of the novel string *jumping jack*. Those prior experiences contradict on many, if not most features, but consistently indicate that *ing* occurs in chunk-final position in the infant's lexicon. Components of the input are only chunked together if they are attended simultaneously, so the ability of prior experience to guide attention is crucial.

As these examples indicate, a synthesis between chunking and memory trace models allows for segmentation to adapt to the characteristics of the native language. Neither chunking nor memory trace models alone are capable of simulating cue-based statistical learning (though as Perruchet and Tillmann, 2010, demonstrate, chunking models can be sensitive to acoustic cues—but in that formulation, Parser has no way of learning cues), whereas in combination they can do so. Early in learning, chunking proceeds largely in the absence of strategic or cue-based learning. However, the candidate words that are discovered via chunking contain potential acoustic cues to word boundaries. These chunks (with embedded cues) are fed to long-term memory, allowing for detection of similarity across chunks. To the extent that these chunks correspond to words, and to the extent that there are acoustic cues correlated with word boundaries, memory trace models are capable of discovering the acoustic features of the input that predict word boundaries. Then these features can be used to bias attention in the chunking process, leading to subsequent chunkings that are more likely to match the acoustic characteristics of prior words (e.g., stress or phonotactic patterns). Although there is much computational work to be done to flesh out this framework, the combination of chunking and memory trace models presents the potential to incorporate a far wider range of statistical learning phenomena than any prior model.

Additional Benefits of a Synthesized Framework

In addition to providing a mechanistic account for cue-based statistical learning, the synthesis of chunking and exemplar memory models has a secondary benefit: It may bring chunking models into better alignment with human performance in statistical learning tasks. Although chunking models have been largely successful in simulating conditional statistical learning, there are some phenomena that present difficulties for chunking models. Pairing chunking with long-term memory processes may alleviate some of these difficulties. Without a fully specified computational model, it is impossible to exhaustively demonstrate an advantage for this framework in comparison to chunking models in isolation. Nevertheless, there are conditional statistical learning phenomena that suggest our framework may be a better fit to human performance than chunking models in isolation.

For example, one conditional learning phenomenon that presents difficulty for chunking in isolation is the endorsement of illusory words. The compelling nature of illusory words was demonstrated by Endress and Mehler (2009), who familiarized participants with a language containing trisyllabic words that were each generated from a "prototype" word from which they differed by a single syllable (e.g., the prototype *kofuta* might spawn the words *kobita*, *lifuta*, and *kofuno*). The prototype word was never presented in the language; only the subsidiary words that differed from the prototype were presented to participants. After exposure, participants were able to distinguish words they heard from foil items with low-transitional probabilities. However, they selected at chance between words that actually occurred in the exposure and the prototypical words that they had never heard. This suggests that exposure to the exemplars (clustered around the prototype in similarity space) enabled participants to form a representation of the prototype even though they never heard it (e.g., Bomba & Siqueland, 1983).

Parser does not account for prototype formation. But exemplar memory models are perfectly suited to capture this phenomenon. For example, in MINERVA 2 (Hintzman, 1984) the words from the language would be represented as feature vectors. Assuming that the words have been segmented successfully from the language—due, from our perspective, to the process of chunking—any test item that overlaps with the feature vectors will activate those memories. Because all of the words generated from the prototype are equally similar to the prototype (all differ from it by one syllable), any test trial on which the prototype occurs will activate all of the prototype-generated words equally. The memory trace that is returned from a probe of the prototype will sum across all of the prototype-generated words. Information that is inconsistent across the words (the single syllable changed from the prototype) will be inconsistent across these memory traces, and thus cancel out. But the information that is consistent—the components of the words that are consistent with the prototype—will be strongly active. In this way, exemplar memory models react as though they have previously experienced the prototype, even though it has never been presented. Unlike naïve chunking models, chunking models that feed into an exemplar memory system that is sensitive to similarity can account for verification of "illusory" prototype words.

A similar problem (for accounts of statistical learning in which chunking operates in isolation) is presented by evidence that learners are able to detect conditional regularities across category members, rather than individual exemplars. The most striking demonstration of this is an experiment by Brady and Oliva (2008), in which participants saw a series of images (presented sequentially) with strong conditional relations between certain *categories* of images; for example, the category kitchen might predict the category office. Importantly, each exemplar of a category (i.e., each unique image) was only seen once. Chunking models such as Parser (Perruchet & Vinter, 1998) have no natural way to identify the category-level conditional relations, because these models are extracting and storing previously seen exemplars. Because no exemplar is ever reexperienced, none of the stored chunks would

be informative. However, the process of integration involves comparison across exemplars, highlighting consistent information (in this case, category membership). Once this comparison occurs and learners discover the category-level regularity, it is possible to identify the conditional relation among the categories.

Summary

In isolation, neither the process of extraction nor the process of integration can provide a complete account of statistical learning. Without sensitivity to similarity, and the ability to compare across exemplars, extraction provides no explanation of how categorical structure is learned. Conversely, integration provides no explanation of how exemplars are segmented from the input, and models of integration—such as memory trace models (e.g., Hintzman, 1984; McClelland & Rumelhart, 1985)—often assume that the input for learning has previously been segmented by some other process (e.g., P. W. Jusczyk, 1993). The partial nature of each solution is illustrated by the combination of chunking and memory trace models we have chosen to instantiate extraction and integration. Separately, trace memory models and chunking models can each account for different aspects of statistical learning. Both kinds of models are based on general principles of memory, but neither model alone is complete.

The benefit of combining the processes of extraction and integration in a single approach is straightforward. Unlike prior accounts, the framework we have proposed can account for all three aspects of statistical learning: conditional, distributional, and cue based. Of course, integrating these processes into a computational model (as opposed to a verbal framework) presents a number of computational challenges. First, information must be encoded in some format that is sensitive to similarity. This is standard in memory trace approaches but would require changes to many extant models of extraction, such as chunking models. Second, the distributional characteristics of chunks stored in long-term memory must serve to guide subsequent extraction. In the combination of exemplar memory models and chunking that we have outlined, this would be done because prior experience would bias attention during the formation of chunks. On this account, conditional statistics are detected via chunking, mediated by attention and working memory limitations. Distributional statistics arise due to the accretion of exemplars in long-term memory, and the distributional characteristics of those exemplars. Cue-based statistical learning occurs when the distributional characteristics of the chunks in long-term memory influences attention during subsequent chunking.

This account is not fully specified, but it possesses a number of advantages even so. First and foremost, our framework unifies sensitivity to conditional statistics with distributional and cue-based statistical learning. As such, it encompasses a wider range of phenomena than most prior models of statistical learning, which have primarily been limited to conditional statistical learning. This approach broadens statistical learning beyond word segmentation and suggests ways in which modeling of statistical learning can incorporate phenomena such as syntactic learning. Our mechanism of comparing across prior memory traces has natural connections to previous work in syntactic learning involving item-based or frame-based learning (e.g., MacWhinney, 1982; Mintz, 2003). Second, this framework is based on the characteristics of human memory. This potentially provides an avenue to integrate the literature on statistical learning with the literature on implicit learning more generally. Finally, the framework that we have proposed in this review is consistent with compelling evidence that statistical learning results in discrete, chunklike representations (e.g., Graf Estes et al., 2007; Orbán et al., 2008).

Conclusion

The benefit of any theoretical account is not only its ability to account for previously observed data in a different way than previous accounts but also its ability to make novel predictions. The extraction and integration framework we have outlined does so in a variety of different domains. Here, we highlight one particular domain—word segmentation—that has been particularly important in research on statistical learning. Doing so will illuminate how the extraction and integration framework incorporates a wider range of data than previous accounts, does so in a unique manner, and yields new predictions about statistical learning. As before, we describe this framework using a chunking model and an exemplar memory model as the specific computational implementations of extraction and integration.

How Development and Experience Alter Word Segmentation

From our perspective, sensitivity to both statistical and phonological cues is due to intimately related processes. Previously identified chunks (extracted via conditional statistical learning) provide the opportunity to discover acoustic regularities (identified via distributional statistical learning) that can alter subsequent learning. This suggests that the very earliest form of word segmentation is due to universal cues, such as those provided by utterance boundaries (e.g., Seidl & Johnson, 2006) and statistical coherence in the speech stream (e.g., Thiessen & Saffran, 2003). But learning should quickly adapt to the characteristics of the native language as infants discover word forms that provide them with information about the acoustic regularities in their linguistic environment (Thiessen & Saffran, 2007).

Note that from this perspective, the central mechanisms of statistical learning do not change across development. Developmental differences in statistical learning (e.g., Howard & Howard, 2001; Hudson Kam & Newport, 2005) are due to changes in factors that are peripheral to these mechanisms (Thompson-Schill, Ramscar, & Chrysikou, 2009). Many of these factors are altered by both maturation and prior experience. For example, even very young infants are sensitive to the distinction between stressed and unstressed syllables (e.g., P. W. Jusczyk & Thompson, 1978). But it is only after extensive experience with English that infants begin to use stress as a cue to word segmentation (e.g., Jusczyk, Houston, & Newsome, 1999). Curtin et al. (2005) argue that this is due to the fact that infants' experience with the language makes stress a much more salient part of their representation of speech. From our perspective, prior experience does not exert a direct effect on statistical learning, because experience is not a mechanism in and of itself. Instead, prior experience influences a learner's perception and attention in ways that determine which aspects of the stimulus are represented in working memory and long-term memory.

The hypothesis that word segmentation begins with the ability to chunk items from the input suggests that word segmentation

should begin early in life, as the ability to chunk is presumably an early developing one depending largely on memory. This is inconsistent with the common citation of 7 months as the earliest age at which infants are capable of segmenting words from fluent speech (e.g., E. K. Johnson & Seidl, 2008; P. W. Jusczyk & Aslin, 1995). It may be that this inconsistency can be resolved. There is evidence that word segmentation begins much earlier than 7 months, at least in some situations. The Jusczyk and Aslin experiments likely underestimated infants' word segmentation abilities. In those experiments, children were exposed to only 12 tokens of each word. In subsequent experiments, including some that have demonstrated segmentation earlier than 7 months of age (e.g., Thiessen & Saffran, 2003), infants have been exposed to many more tokens of each word, which facilitates segmentation (e.g., Saffran et al., 1996; Thiessen, Hill, & Saffran, 2005). As such, the question of when infants first begin to be able to segment words from fluent speech remains open.

Moreover, when presented with input where the chunking problem is simplified, infants may be able to succeed at a very young age. The first simplification that can help infants chunk is the presence of pauses at the beginning and ending of utterances (e.g., Seidl & Johnson, 2006). These pauses serve as a natural indication of the beginning and end of a chunk, which increases the likelihood that infants' early chunks will have at least some correspondence with word boundaries. The second simplification that can help infants chunk is an unambiguous statistical structure. In natural languages, occurrences of words are often widely spaced, and perceptual primitives occur in many different words. Both of these factors increase the difficulty of finding wordlike chunks in the input (e.g., Perruchet & Vintner, 1998). When presented with simplified input, infants are capable of benefiting from statistical structure at a much younger age (e.g., Kirkham et al., 2002). Thus, the use of chunking as a mechanism of extraction predicts that infants can segment words from fluent speech from a very age, so long as they are provided with simplified input. Note that both of these simplifications are available cross-linguistically and that infants can benefit from them without any prior knowledge of the structure of the language. If this account is correct, even very young infants should show some ability to segment lexical forms from fluent speech if the input has appropriate pauses (e.g., Seidl & Johnson, 2006) or a clear statistical structure (e.g., Onnis, Christiansen, Chater, & Gómez, 2003).

Early in the process of extracting a lexicon from the input, then, infants depend on cues that are available cross-linguistically: pauses, words presented in isolation, and unambiguous statistical structure. The words that they discover from these early cues may provide an important database from which to extract linguistic regularities relevant to word segmentation (e.g., Thiessen & Saffran, 2003). Likely the first piece of information infants can use is the presence of familiar words in the speech stream (e.g., Bortfeld et al., 2005). Previously known chunks provide an anchor for chunking subsequent information in much the same way as utterance-initial pauses (Perruchet et al., 2004). That is, for an infant who already knows *baby*, the phrase "baby bottle" now presents a much easier segmentation than it does for an infant who has previously chunked neither word. Additionally, experience with the language provides an infant with information about many other language-specific acoustic cues to word boundaries, including stress (e.g., E. K. Johnson & Jusczyk, 2001) and phonotactics (Mattys, Jusczyk, Luce, & Morgan, 1999). This adaptation to the native language may explain why younger infants fail to segment words from fluent speech when given the same amount of input as older infants (e.g., P. W. Jusczyk & Aslin, 1995).

The effect of prior experience on attention is applicable to a wide variety of cues, including lexical stress and phonotactics. If one of the sequences in the speech stream matches a previously experienced chunk, then that memory is activated most strongly and guides segmentation (e.g., Bortfeld et al., 2005). If, however, the speech stream contains novel words, then no single prior memory trace is strongly activated. Instead, multiple traces are activated to a degree depending on their similarity to the input. Thus, the features that are consistent across the majority of previous words come to guide segmentation (e.g., Hintzman, 1986; Thiessen & Pavlik, in press). For infants learning English, the majority of words they know are likely to be stressed on their first syllable (e.g., Cutler & Carter, 1987). Novel incoming speech activates many or all of these words, depending on the similarity of the input to prior chunks (e.g., Goldinger, 1998). Elements of those words that conflict cancel each other out, and only the information that is consistent across words is returned from long-term memory (e.g., Hintzman, 1984). In this case, novel incoming speech activates the knowledge that words are stressed on their first syllable, and this serves to guide subsequent chunking behavior. Attention is distributed across time in such a way that stressed syllables mark the onset of candidate chunkings.

Increased native-language experience typically means that older infants will be more successful in segmenting fluent speech than younger learners (e.g., P. W. Jusczyk & Aslin, 1995). However, this is not always the case. When infants are presented with speech that violates the regularities they have learned from their prior experience, older infants will perform worse than younger infants (e.g., Thiessen & Saffran, 2003). This is because older infants are reliant on the similarity structure in previously identified lexical forms and will mis-segment the input. For example, 9-month-olds presented with a stream of speech in which words are segmented on the second syllable (as in the phrase *guiTAR is*) will mistakenly treat the stressed syllable as the onset of a chunk (e.g., E. K. Johnson & Jusczyk, 2001). Younger infants, who have not yet identified the acoustic regularity, will chunk randomly and eventually identify the statistical structure of the input (e.g., Thiessen & Saffran, 2003). As such, familiarity with the acoustic regularities in lexical forms is a double-edged sword. It allows for faster, more efficient learning in input where the input follows the regularities that the infant has previously learned. Instead of chunking randomly, the infant's first candidate segmentations of the stream follow the regularities of the language, and are thus more likely to be correct. However, when the learner is placed in an environment that violates the regularities they have learned, learning will be slow or inaccurate (e.g., Finn & Hudson, 2008).

This proposed account has the advantage of integrating several phenomena previously seen as distinct. Many accounts of word segmentation have treated acoustic regularities as being in competition with statistical regularities, such that infants might choose to focus on one or the other (e.g., E. K. Johnson & Seidl, 2008). Similarly, some theories have suggested that words in isolation might be sufficient for identifying a lexicon or for identifying regularities among words in isolation, such that statistical learning (read as sensitivity to transitional probabilities) is never necessary

for learning in the infants' natural environment (e.g., Brent & Siskind, 2001; E. K. Johnson & Jusczyk, 2001). From our perspective, all of these phenomena arise from the operation of the same central processes. Words in isolation are preferentially chunked, because pauses provide a strong cue to group elements within (and not across) pause boundaries. Sensitivity to regularities such as stress arises due to the similarity structure of previous chunks influencing subsequent chunking. And sensitivity to the statistical structure measured by transitional probabilities arises due to ubiquitous processes of decay and interference in chunking (e.g., Perruchet & Vinter, 1998).

Comparison to Other Accounts

The most obvious distinction between our framework and other models of statistical learning is that most other theoretical accounts of statistical learning have solely attempted to explain conditional statistical learning (e.g., Frank et al., 2010; Gambell & Yang, 2004; Perruchet & Vinter, 1998). Our framework is different from these in that in addition to accounting for the extraction (conditional statistical learning), it encompasses integration (distributional statistical learning). Even if we limit the scope of comparison to conditional statistical learning, however, our framework differs from many other accounts. Our framework falls squarely within the tradition of clustering models of statistical learning, because one of the core principles of the account is that learners are extracting and storing discrete representations (e.g., words) from continuous input. Most connectionist models of word segmentation, by contrast, can be characterized as boundary-finding models (e.g., Christiansen et al., 1998; Gambell & Yang, 2004). These models learn to predict the next element in a sequence on the basis of previous elements (e.g., Elman, 1990). Word boundaries are identified as regions where the next prediction is poor. It should be noted, though, that although many connectionist networks are boundary finding, this does not mean that all connectionist networks are boundary finding. It is possible to create a connectionist network that extracts word forms or even that represents discrete units in much the same way as chunking models (e.g., Boucher & Dienes, 2003).

A second important contrast between this framework and many (though certainly not all) models of conditional statistical learning is that this framework does not rely on the calculation of transitional probabilities for segmentation. From our perspective, transitional probabilities are a useful descriptor of the statistical structure of the input, but they do not guide learning. Instead, learning occurs due to the competition between potential groups (e.g., competition between words and spurious groupings of syllables that occur across word boundaries) in the process of extraction. This allows factors other than conditional relations, especially perceptual cues to grouping, to influence which items are extracted from continuous input. Transitional probabilities are incapable of incorporating other sources of information; by definition, they are only sensitive to the conditional relations between syllables (e.g., Aslin et al., 1998). As such, theories that seek to incorporate transitional probabilities and phonological cues to segmentation tend to do so in a stagelike manner, where phonological cues are used in one stage and transitional probabilities in another (e.g., Mersad & Nazzi, 2011; Shukla et al., 2007). Our framework, by contrast, suggests that sensitivity to conditional and perceptual cues are incorporated into the same process of extraction.

Beyond the characterization of the processes underlying conditional statistical learning, the framework outlined in this review is distinct in that it attempts to encompass a wider range of statistical learning phenomena than prior accounts: not just sensitivity to conditional probabilities, or exemplar distribution, or cue learning, but all three. In this regard, the most similar model to the account we have laid out in this review is StaGe model (Adriaans & Kager, 2010). The StaGe model invokes two processes: sensitivity to conditional probabilities (implemented via observed/expected probabilities, a statistic closely related to mutual information) and generalization. Generalization allows the model to make predictions about segmentation of novel words on the basis of phonotactic patterns in the lexicon. These generalizations are governed by principles derived from optimality theory, such that the likeliest generalization (absent other evidence) is the one that violates the fewest constraints. In this way, the model uses the words that it segments (via probability) as a basis for generalizing about the phonotactic structure of the input.

Despite the similarity in scope between the StaGe (Adriaans & Kager, 2010) model and our perspective, there are two important differences. First, the StaGe model uses a boundary-finding model to achieve segmentation (via observed/expected probabilities), whereas our framework depends on a clustering approach, one that extracts candidate word forms into a proto-lexicon. Second, the StaGe model achieves generalization through reference to optimality theory (e.g., Prince & Smolensky, 1997). This has the advantage of being a more formal—and thus better constrained—proposal than our use of similarity across exemplars. However, it also renders the StaGe model explicitly linguistic: The processes that govern generalization in language should be different than the processes that govern generalization for nonlinguistic stimuli. By contrast, though similarity is not operationalized in our proposal, our perspective suggests that the same process should govern generalization across domains: comparing the present instance with prior exemplars on the basis of similarity. This is, of course, a distinction that is susceptible to empirical testing. Indeed, recent results suggest that at least some forms of linguistic generalizations are due to domain-general processes of similarity, rather than domain-specific constraints (e.g., Thiessen, 2011a).

Novel Predictions and Next Steps

By combining conditional and distributional learning, the extraction and integration framework accounts for a wider range of statistical learning phenomena than prior accounts. To do so, the framework invokes a set of processes and specifies the way they interact. In addition to providing a new explanatory framework for statistical learning, this account leads to a set of novel predictions. Identifying some of these predictions will help to better explicate the framework as well as indicate some of the ways in which it might be falsifiable. To do so, we discuss two sets of predictions: one relating to the role of attention in individual differences in conditional statistical learning and the other relating to the discovery of prototypical elements among sequentially presented stimuli.

Although statistical learning is often described as a kind of incidental or implicit learning (e.g., Perruchet & Pacton, 2006; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), this should

not be taken to mean that statistical learning can proceed in the absence of attention. A variety of converging evidence indicates that attention is important, even necessary, for statistical learning. Infants identify statistical structure more successfully in stimuli to which they are more likely to attend (Thiessen et al., 2005). Baker et al.z(2004) found that adults presented with visual displays were only able to detect conditional relations among elements to which they attended. Similarly, when adults are distracted by a secondary task, they fail to detect the conditional relations in a to-be-segmented language (Toro, Sinnett, & Soto-Faraco, 2005). Even though statistical learning occurs in the absence of learners' intent to discover statistical patterns (e.g., Fletcher et al., 2005), some degree of attention to the input is necessary for learning.

Our framework provides a natural way to account for this effect of attention. From our perspective, consistent with chunking models (e.g., Perruchet & Vinter, 1998), the process of extraction involves the binding together of elements of the input into a discrete representation. That cannot occur unless the elements are held simultaneously in attention. This leads to a series of predictions about the process of conditional statistical learning. One is that conditional relations involving more salient elements in the input—the elements that are most likely to attract attention —should be learned more quickly than conditional relations involving less salient elements. This prediction is consistent with anecdotal observations about conditional statistical learning (e.g., Hayes & Clark, 1970) but has not been extensively tested. A second prediction is that attention may serve to mediate individual differences in statistical learning. Learners who are better able to maintain attention and resist distraction should be able to extract items (such as words) from the input more successfully than learners who have more difficulty maintaining attention. Individual differences in attention are likely to have the largest effect in childhood, when executive control is developing, leading to relatively larger individual differences. Although these predictions have yet to be assessed, the extraction and integration framework clearly suggests that the items learners extract from the speech stream may not perfectly reflect the statistical structure of the input; instead, extraction is influenced by extra-statistical factors such as attention.

In much the same way that the extraction and integration framework predicts that learners' representations may not perfectly reflect the conditional statistical structure of the input, the framework predicts that it should be possible to observe distortions of the distributional statistical structure of the input. This prediction arises from the framework's method of combining conditional and distributional statistics: Distributional statistics are computed over the items extracted from the input. If an item is never extracted from the input, it will not influence the learner's emerging sense of the central tendency and variation of the distributional structure of the input. That is, learners may miss some aspect of the distributional structure of the input by failing to extract a portion of the input. If this is the case, it should be possible to influence a learner's representation of distributional information by making some items easier to extract, and other items more difficult.

As an example, consider Endress and Mehler's (2009) demonstration that exposure to a set of words leads to the representation of a "prototype" word that has never been seen. After exposure to words like *kobita*, *lifuta*, and *kofuno*, participants endorse *kofuta* as familiar, even though they have not previously heard it. The extraction and integration framework predicts that the endorsement of this prototype is critically dependent on the extraction of the individual words. Only when the words have been extracted can learners compare across them, and integrate their information in such a way that the central tendency (i.e., the prototype word) becomes apparent. One way to affect the ease with which these words are extracted would be to insert pauses within words (e.g., between the first and second syllable of each word). Another possibility would be to alter the coarticulatory cues such that the first and second syllables of each word sound as though they are from different utterances. Either kind of acoustic cue would lead learners to segment part-words (such as *bitali* or *futako*), rather than words, from the speech stream (e.g., E. K. Johnson & Jusczyk, 2001). We predict that this would inhibit the representation of a prototype word, because the items that learners would integrate across (the part-words they have extracted) do not have a consistent prototypical structure. Importantly, this result would demonstrate a difference even though the statistical structure of the input is identical to the statistical structure of input without acoustic cues. Such a result would be consistent with our framework's prediction that rather than creating a prototypical representation by directly accessing the statistical structure of the input, the prototype is critically dependent on the items learners extract from the input.

Although we have primarily focused on how this framework relates to statistical learning in the domain of word segmentation, it is intended to apply much more broadly. One natural, linguistically relevant extension of this framework is to assess its fit to learning syntactic regularities. One critique of statistical learning approaches to language is that statistical learning may be unable to account for learning of syntactic patterns (e.g., Marcus, 2000; Marcus & Berent, 2003). For example, although the relevant units for word segmentation are speech sounds that are directly perceptible in the input, the relevant units for discovering syntactic regularities are categories—such as *noun* and *verb*—that are not directly available. As such, any statistical learning approach to syntactic learning must incorporate a mechanism for learning about these categories from the exemplars (i.e., word forms) in the input, which should tap into the same processes that we have termed distributional statistical learning. Learning of syntactic regularities is a domain in which sensitivity to conditional and distributional statistical learning need to be examined in concert (e.g., Thompson & Newport, 2007), and therefore provides an intriguing test of the principles underlying the extraction and integration framework. For example, one prediction of this framework is that the same mechanisms that underlie word segmentation underlie syntactic learning, and therefore learning at one level of language should influence learning at other levels (e.g., Onnis & Thiessen, in press).

Summary and Conclusion

The goal of this framework is twofold: first, to outline the range of statistical learning phenomena beyond conditional statistical learning and, second, to explore the possibility of synthesizing all of these phenomena into a unified account. With respect to the first goal, we have suggested two additional kinds of accomplishments to which the term *statistical learning* is routinely applied. One is learning based on a distribution of exemplars (e.g., Maye et al.,

2002; Thiessen, 2011b; Thiessen & Yee, 2010). The other is learning to identify perceptual cues to aspects of the statistical structure of the input that are not directly perceptible, such as word boundaries (e.g., Saffran & Thiessen, 2003; Thiessen & Saffran, 2007). These different aspects of statistical learning have rarely been considered in totality. Most modeling and theorizing about statistical learning has been focused on conditional statistical learning. Importantly, our framework posits that these types of learning are linked by more than a label. We believe that a satisfactory approach to statistical learning should integrate all three aspects of statistical learning. Doing so enriches our understanding of the processes underlying statistical learning.

We propose that a complete account of statistical learning must incorporate two interdependent processes: one that extracts statistically coherent items from the input and one that compares and integrates information across those items. The extracting process (which we have described in terms of chunking) is responsible for sensitivity to conditional relations in the input. The process of integration (which we have described in terms of an exemplar memory model) enables sensitivity to distributional information. Additionally, the process of comparison allows for learning of cues related to statistical structure (such as cues to word boundary), allowing the process of extraction to adapt to the characteristics of the input (e.g., the phonological characteristics of a native language). These two processes map onto different aspects of memory. Chunking is dependent on working memory, whereas comparison across previously extracted chunks depends on long-term memory. Unlike most prior models of statistical learning, which are focused solely on conditional statistical learning (e.g., Christiansen et al., 1998; Frank et al., 2010), our framework incorporates both forms of human memory, allowing it to encompass a much wider range of statistical learning phenomena.

In the past decade, research on statistical learning has expanded dramatically from its original focus on word segmentation. Despite this expansion, however, virtually all of the models advanced to explain statistical learning have focused on conditional statistical learning (e.g., Goldwater et al., 2009; Orbán et al., 2008; Perruchet & Vinter, 1998). The extraction and integration framework moves beyond the focus on conditional statistical learning by combining conditional statistical learning with distributional statistical learning. From this perspective, sensitivity to statistical structure in the input arises from processes that are integral to memory such as decay, activation, and interference. Although this framework is not a fully specified computational model, and developing such a model is an important challenge, it does have a set of important strengths. It encompasses a wider range of statistical learning phenomena than previous accounts. It provides a framework for statistical learning that does not require explicit computation of conditional or distributional statistics, rendering it psychologically plausible and consistent with accounts of implicit learning more generally (e.g., Boucher & Dienes, 2003; Reber, 1967). Finally, it provides a way of describing how learners move from a reliance on statistical structure to adapting to the perceptual characteristics of the input, such as the discovery of acoustic cues to word segmentation.

References

Adriaans, F. C., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language, 62,* 311–331. doi:10.1016/j.jml.2009.11.007

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistic by 8-month-old infants. *Psychological Science, 9,* 321–324. doi:10.1111/1467-9280.00063

Baker, C. I., Olson, C. R., & Behrmann, M. (2004). Role of attention and perceptual grouping in visual statistical learning. *Psychological Science, 15,* 460–466. doi:10.1111/j.0956-7976.2004.00702.x

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition, 106,* 1382–1407. doi:10.1016/j.cognition.2007.07.005

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

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science, 16,* 298–304. doi:10.1111/j.0956-7976.2005.01531.x

Boucher, L., & Dienes, Z. (2003). Two ways of learning associations. *Cognitive Science, 27,* 807–842. doi:10.1207/s15516709cog2706_1

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes: Extracting categorical regularities without conscious intent. *Psychological Science, 19,* 678–685.

Brent, M. R. (1999). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences, 3,* 294–301. doi:10.1016/S1364-6613(99)01350-9

Brent, M. R., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81, B33–B44*. doi:10.1016/S0010-0277(01)00122-6

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81. doi:10.1016/0010-0285(73)90004-2

Chomsky, N. (1980). *Rules and representations*. New York, NY: Columbia University Press. doi:10. 1017/S0140525X00001515

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York, NY: Harper & Row.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13,* 221–268. doi:10.1080/016909698386528

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

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities. *Psychological Science, 17,* 905–912. doi:10.1111/j.1467-9280.2006.01801.x

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,* 1119–1130. doi:10.1037/0278-7393.30.5.1119

Curtin, S., Mintz, T. H., & Christiansen, M. H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition, 96,* 233–262. doi:10.1016/j.cognition.2004.08.005

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language, 2,* 133–142. doi:10.1016/0885-2308(87)90004-0

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 113–121. doi:10.1037/0096-1523.14.1.113

Dougherty, T. M., & Haith, M. M. (2002). Infants' use of constraints to speed information processing and to anticipate events. *Infancy, 3,* 457–473. doi:10.1207/S15327078IN0304_03

Edwards, C. A., Jagielo, J. A., Zentall, T. R., & Hogan, D. E. (1982). Acquired equivalence and distinctiveness in matching to sample by pigeons: Mediation by reinforcer-specific expectancies. *Journal of Experimental Psychology: Animal Behavior Processes, 8,* 244–259. doi:10.1037/0097-7403.8.3.244

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179–211. doi:10.1207/s15516709cog1402_1

Emberson, L., Liu, R., & Zevin, J. D. (2009, August). *Statistics all the way down: How is statistical learning accomplished using varying productions of novel, complex sound categories?* Paper presented at the 31st annual meeting of the Cognitive Science Society, Amsterdam, the Netherlands.

Endress, A. D., & Mehler, K. (2009). Primitive computations in speech processing. *The Quarterly Journal of Experimental Psychology, 62,* 2187–2209. doi:10.1080/17470210902783646

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116,* 752–782. doi:10.1037/a0017196

Finn, A., & Hudson, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition, 108,* 477–499. doi:10.1016/j.cognition.2008.04.002

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 99,* 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,* 521–537. doi:10.1037/0096-3445.134.4.521

Fletcher, R. C., Zafiris, O., Frith, C. D., Honey, R. A. E., Corlett, R., Zilles, K., & Fink, G. R. (2005). On the benefits of not trying: Brain activity and connectivity reflecting the interactions of explicit and implicit sequence learning. *Cerebral Cortex, 15,* 1002–1015. doi:10.1093/cercor/bhh201

Frank, M. C., Goldwater, S., Griffiths, T., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition, 117,* 107–125. doi:10.1016/j.cognition.2010.07.005

Gambell, T., & Yang, C. (2004). *Statistics learning and universal grammar: Modeling word segmentation*. Paper presented at the 20th International Conference on Computational Linguistics, Geneva, Switzerland.

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science, 33,* 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,* 251–279. doi:10.1037/0033-295X.105.2.251

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112,* 21–54. doi:10.1016/j.cognition.2009.03.008

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13,* 431–436. doi:10.1111/1467-9280.00476

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007).). Can infants map meaning to newly segmented words? Statistical segmentation and word leaning. *Psychological Science, 18,* 254–260. doi:10.1111/j.1467-9280.2007.01885.x

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General, 121,* 480–506. doi:10.1037/0096-3445.121.4.480

Griffiths, O., & Mitchell, C. J. (2008). Negative priming reduces affective ratings. *Cognition & Emotion, 22,* 1119–1129. doi:10.1080/02699930701664930

Haith, M. M., Wentworth, N., & Canfield, R. L. (1993). The formation of expectations in early infancy. In C. Rovee-Collier & L. P. Lipsitt (Eds.), *Advances in infancy research*. Norwood, NJ: Ablex.

Hall, G. (1991). *Perceptual and associative learning*. Oxford, England: Clarendon Press.

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review, 56,* 51–65. doi:10.1037/h0062474

Harris, Z. (1955). From phoneme to morpheme. *Language, 31,* 190–222. doi:10.2307/411036

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, NY: Wiley.

Hintzman, D. L. (1976). Repetition and memory. *Psychology of Learning and Motivation, 10,* 47–91. doi:10.1016/S0079-7421(08)60464-8

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, 16,* 96–101. doi:10.3758/BF03202365

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428. doi:10.1037/0033-295X.93.4.411

Hintzman, D. L., Block, R. A., & Summers, J. J. (1973). Contextual associations and memory for serial position. *Journal of Experimental Psychology, 97,* 220–229. doi:10.1037/h0033884

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of Acoustical Society of America, 119,* 3059–3071. doi:10.1121/1.2188377

Honey, R. C., & Hall, G. (1989). Acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Process, 15,* 338–346. doi:10.1037/0097-7403.15.4.338

Howard, D. V., & Howard, J. H. (2001). When it does hurt to try: Adult age differences in the effects of instructions on implicit pattern learning. *Psychonomic Bulletin & Review, 8,* 798–805. doi:10.3758/BF03196220

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1,* 151–195. doi:10.1080/15475441.2005.9684215

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

James, W. (1890). *The principles of psychology*. New York, NY: Holt. doi:10. 1037/11059-000

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,* 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,* 440–455. doi:10.1080/15250000802329321

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

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29,* 1–23. doi:10.1006/cogp.1995.1010

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development, 64,* 675–687. doi:10.2307/1131210

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language, 32,* 402–420. doi:10.1006/jmla.1993.1022

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39,* 159–207. doi:10.1006/cogp.1999.0716

Jusczyk, P. W., & Thompson, E. (1978). Perception of a phonetic contrast in multisyllabic utterances by 2-month-old infants. *Attention, Perception, & Psychophysics, 23,* 105–109. doi:10.3758/BF03208289

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83,* B35–B42. doi:10.1016/S0010-0277(02)00004-5

Kruschke, J. K. (2001). Towards a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45,* 812–863. doi:10.1006/jmps.2000.1354

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General, 124,* 161–180. doi:10.1037/0096-3445.124.2.161

Lehle, C., & Hubner, R. (2008). On-the-fly adaptation of selectivity in the flanker task. *Psychonomic Bulletin & Review,15,* 814–818. doi:10.3758/PBR.15.4.814

Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition, 122,* 241–246.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54,* 358–368. doi:10.1037/h0044417

Lidz, J., Gleitman, H., & Gleitman, L. (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition, 87,* 151–178. doi:10.1016/S0010-0277(02)00230-5

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail. *Journal of Acoustical Society of America, 102,* 1134–1140. doi:10.1121/1.419865

MacWhinney, B. (1982). Basic syntactic processes. In S. Kuczaj (Ed.), *Language development, 1: Syntax and semantics* (pp. 73–136). Hillsdale, NJ: Erlbaum.

Madole, K. L., & Oakes, L. M. (1999). Making sense of infant categorization: Stable processes and changing representations. *Developmental Review, 19,* 263–296.

Marcus, G. F. (2000). Pabiku and Ga Ti Ga: Two mechanisms infants use to learn about the world. *Current Directions in Psychological Science, 9,* 145–147. doi:10. 1111/1467-8721. 00080

Marcus, G. F., & Berent, I. (2003). Are there limits to statistical learning? *Science, 300,* 53–55. doi:10.1126/science.300.5616.53

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38,* 465–494. doi:10.1006/cogp.1999.0721

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science, 32,* 543–562. doi:10.1080/03640210802035357

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science, 11,* 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,* B101–B111. doi:10.1016/S0010-0277(01)00157-3

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188. doi:10.1037/0096-3445.114.2.159

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86,* B33–B42. doi:10.1016/S0010-0277(02)00157-9

Mersad, K., & Nazzi, T. (2011). Transitional probabilities and positional frequency phonotactics in a hierarchical model of speech segmentation. *Memory & Cognition, 39,* 1085–1093.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46,* 505–512. doi:10.3758/BF03208147

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90,* 91–117. doi:10.1016/S0010-0277(03)00140-9

Mintz, T. H. (2006). Finding the verbs: Distributional cues to categories available to young learners. In K. Hirsh-Pasek & R. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 31–63). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195170009.003.0002

Mirman, D., Magnuson, J. S., Graf Estes, K., & Dixon, J. A. (2008). The link between statistical segmentation and word learning in adults. *Cognition, 108,* 271–280. doi:10.1016/j.cognition.2008.02.003

Newport, E. L., & Aslin, R. N. (2004). Learning at distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48,* 127–162. doi:10.1016/S0010-0285(03)00128-2

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 924–940. doi:10.1037/0278-7393.28.5.924

Onnis, L., Christiansen, M., Chater, N., & Gómez, R. (2003). *Reduction of uncertainty in human sequential learning: Evidence from artificial language learning*. Paper presented at the 25th Annual Conference of the Cognitive Science Society, Boston, Massachusetts.

Onnis, L., & Thiessen, E. D. (in press). *Language experience changes subsequent learning*. Cognition.

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,* 2745–2750. doi:10.1073/pnas.0708424105

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development, 80,* 674–685. doi:10.1111/j.1467-8624.2009.01290.x

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition, 36,* 1299–1305. doi:10.3758/MC.36.7.1299

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences, 10,* 233–238. doi:10.1016/j.tics.2006.03.006

Perruchet, P., & Tillmann, B. (2010). Exploiting multiple sources of information in learning an artificial language: Human data and modeling. *Cognitive Science, 34,* 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,* 573–583. doi:10.1037/0096-3445.133.4.573

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39,* 246–263. doi:10.1006/jmla.1998.2576

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15,* 285–290. doi:10.3758/BF03213946

Prince, A., & Smolensky, P. (1997, March 14). Optimality: From neural networks to universal grammar. *Science, 275,* 1604–1610. doi:10.1126/science.275.5306.1604

Rakison, D. H. (2004). Infants' sensitivity to correlations between static and dynamic features in a category context. *Journal of Experimental Child Psychology, 89,* 1–30. doi:10.1016/j.jecp.2004.06.001

Rakison, D. H. (2005). A secret agent? How infants learn about the identity of objects in a causal scene. *Journal of Experimental Child Psychology, 91,* 271–296. doi:10.1016/j.jecp.2005.03.005

Rakison, D. H., & Lupyan, G. (2008). Developing object concepts in infancy: An associative learning perspective: I. Introduction. *Monographs of the Society for Research in Child Development, 73,* 1–29.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6,* 855–863. doi:10.1016/S0022-5371(67)80149-X

Reber, A. S., & Lewis, S. (1977). Implicit learning: An analysis of the form and structure of a body of tacit knowledge. *Cognition, 5,* 333–361. doi:10.1016/0010-0277(77)90020-8

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22,* 425–469. doi:10.1207/s15516709cog2204_2

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814. doi:10.1037/0278-7393.21.4.803

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition, 81,* 149–169. doi:10.1016/S0010-0277(01)00132-9

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996, December 13). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928. doi:10.1126/science.274.5294.1926

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70,* 27–52. doi:10.1016/S0010-0277(98)00075-4

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8,* 101–105.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology, 39,* 484–494. doi:10.1037/0012-1649.39.3.484

Samuelson, L. K., & Smith, L. B. (2000). Children's attention to rigid and deformable shape in naming and non-naming tasks. *Child Development, 71,* 1555–1570. doi:10.1111/1467-8624.00248

Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science, 9,* 565–573. doi:10.1111/j.1467-7687.2006.00534.x

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology, 54,* 1–32.

Smith, L. B., Jones, S. S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition, 60,* 143–171. doi:10.1016/0010-0277(96)00709-3

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106,* 1558–1568. doi:10.1016/j.cognition.2007.06.010

Sobel, D. M., & Kirkham, N. Z. (2007). Bayes nets and babies: Infants' developing statistical reasoning abilities and their representation of causal knowledge. *Developmental Science, 10,* 298–306. doi:10.1111/j.1467-7687.2007.00589.x

Stadler, M. A. (1992). Statistical structure and implicit serial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 318–327. doi:10.1037/0278-7393.18.2.318

Stager, C. L., & Werker, J. F. (1998). Methodological issues in studying the link between speech perception and word learning. In C. Rovee-Collier, L. P. Lipsitt, & H. Hayne (Series Eds.), *Advances in infancy research* (pp. 237–256). Stamford, CT: Ablex.

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language, 56,* 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, England: Cambridge University Press. doi:10.1017/CBO9780511576164.003

Thiessen, E. D. (2011a). Domain general constraints on statistical learning. *Child Development, 82,* 462–470. doi:10.1111/j.1467-8624.2010.01522.x

Thiessen, E. D. (2011b). When variability matters more than meaning: The effect of lexical forms on use of phonemic contrasts. *Developmental Psychology, 47,* 1448–1458. doi:10.1037/a0024439

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7,* 53–71. doi:10.1207/s15327078in0701_5

Thiessen, E. D., & Pavlik, P. (in press). iMinerva: A mathematical model of distributional statistical learning. *Cognitive Science*.

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,* 706–716. doi:10.1037/0012-1649.39.4.706

Thiessen, E. D., & Saffran, J. R. (2004). Spectral tilt as a cue to word segmentation in infancy and adulthood. *Perception & Psychophysics, 66,* 779–791. doi:10.3758/BF03194972

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development, 3,* 73–100. doi:10.1080/15475440709337001

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,* 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,* 1–42.

Thompson-Schill, S. L., Ramscar, M., & Chrysikou, E. G. (2009). Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Science, 18,* 259–263. doi:10.1111/j.1467-8721.2009.01648.x

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97,* B25–B34.

Toro, J. M., & Trobalón, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception & Psychophysics, 67,* 867–875. doi:10.3758/BF03193539

Trabasso, T., & Bower, G. H. (1968). *Attention in learning: Theory and research*. New York, NY: Wiley.

Turk-Browne, N. B., Isola, P. J., Scholl, B. J., & Treat, T. A. (2008). Multidimensional visual statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 399–407. doi:10.1037/0278-7393.34.2.399

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

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition, 103,* 147–162. doi:10.1016/j.cognition.2006.03.006

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,* 49–63. doi:10.1016/S0163-6383(84)80022-3

Yerkes, R. M. (1943). *Chimpanzees: A laboratory colony*. New Haven, CT: Yale University Press.

Younger, B. A., & Cohen, L. B. (1986). Developmental change in infants' perception of correlations among attributes. *Child Development, 57,* 803–815. doi:10.2307/1130356

Received June 14, 2011
Revision received August 2, 2012
Accepted September 10, 2012 ■