Psych 229:
Language Acquisition

Lecture 4

Rules & Statistics

Seidenberg, MacDonald, & Saffran 2002

Languages exhibit statistical structure—that is, they show inhomogeneities in the distribution of sounds, words, and phrases. The importance of this type of structure in learning a language is a matter of intense debate, and is tackled by Peña *et al.* (*1*) on page 604 of this issue.

The debate originated with Chomsky's 1957 discussion of the sentence "Colorless green ideas sleep furiously" (*2*). This sentence can be immediately recognized as being well formed (compared to "Ideas colorless sleep furiously green") even though in both cases the probability that these words have previously occurred in this order is close to zero. Chomsky concluded that the statistical properties of language are not central to the characterization of linguistic knowledge, an insight that became part of the foundation of modern linguistic theory. Whether statistical properties are important in language acquisition was largely set aside. Instead, research focused on how the child converges on the rules and other components of grammar using a combination of deductive (nonstatistical) reasoning and innate knowledge (*3*).

Recently, there has been a resurgence of interest in statistical learning, with evidence showing that infants and young children incorporate statistical cues when learning about the sounds of a language, vocabulary, and the structures in which words occur (*4–6*). These findings complement evidence from adults demonstrating the use of statistical information in comprehending and producing utterances, suggesting that similar mechanisms may underlie the learning and use of language (*7, 8*).

The purported debate: Statistical learning vs. innate knowledge

Seidenberg, MacDonald, & Saffran 2002

Peña et al. 2002

Perhaps there are both statistical processes (based on frequency and distribution of elements in language) and grammatical processes (for example, learning and using rules). Statistical learning may be limited to simpler problems such as learning the sounds of a language and building a lexicon. In contrast, the complexities of grammar may require other nonstatistical procedures. Thus, it seems that learning grammar begins where statistical learning ends.

In practice, however, it turns out to be difficult to establish a boundary between "grammatical" and "statistical" learning. Any corpus of linguistic stimuli contains a vast array of cues and potential generalizations. Even in carefully designed experiments, conditions intended to isolate grammatical processes may introduce correlated statistical cues that would support performance.

Two previous attempts to isolate a distinct grammatical form of learning (*9, 10*) raised similar concerns. In each case subsequent analyses suggested that the behavior could instead have arisen from statistical regularities that occurred simultaneously with the grammatical patterns (*11–13*).

Seidenberg, MacDonald, & Saffran 2002

Discussions of statistical learning need to consider two questions illustrated by the "colorless" sentence. First, what kinds of statistics are people, particularly infants, capable of computing? As in the "colorless" example, most research has investigated the transition probabilities between words or syllables. The Peña *et al.* study is a welcome step forward insofar as it addresses questions concerning nonadjacent elements (*15*). Adult learners can track various types of statistics, including some second-order probabilities and long-distance dependencies (*14, 16*).

Peña et al. 2002 (adults)

RULES OR STATISTICS IN LANGUAGE LEARNING?

A Familiarization stream

...PURAKIBELIGATAFODUPUFOKITALIDUBERAGATARADUPULIKIBEFOGA...

B	Generalizations	Probability
1.	Words have PU_KI, BE_GA, TA_DU structure (Peña et al.'s rule, AXC)	1.0
2.	Initial syllables begin with a stop consonant	1.0
3.	Final syllables begin with a stop consonant	1.0
4.	Continuant consonants occur word medially	1.0
5.	PU predicts RA	0.33
6.	PU is not followed by BE	1.0
7.	KI predicts BE	0.5
8.	PURA predicts KI	1.0
9.	RAKI predicts BE	0.5
10.	PUBE is not followed by KI	1.0

C		Experiment 1 forced choice: Choice supported by	Percentage chosen
Word	PURAKI	1 (rule), 2, 3, 4, 5, 8	73.3
Part-word	RAKIBE	3, 7, 9	26.7

D		Experiment 2 forced choice: Choice supported by	Percentage chosen
Rule-word	PUBEKI	1 (rule), 2, 3, 4, 8	49.8
Part-word	RAKIBE	3, 6, 7, 9, 10	49.2

But what are the limits on these capacities and do infants have similar capacities?

Seidenberg, MacDonald, & Saffran 2002

However, recent results suggest that statistical patterns that occur in natural languages are acquired more readily than patterns not found in natural languages (*16–18*). Thus constraints on learning may play a positive role in explaining why language learners acquire only some of the many generalizations afforded by natural language.

But over what *types* of information are statistics learned? Are all information types created equal?

Chomsky's example assumes that people are computing statistics at a single level of linguistic structure—between words in the "colorless" example, and between syllables in the Peña *et al.* study.

Colorless green ideas sleep furiously

For example, words fall into general *types*—*green* is a property or adjective, and *sleep* is an action or verb—that exhibit characteristic distributions. The "colorless" sentence conforms to these distributions in English, whereas "Ideas colorless sleep furiously green" does not (*19*).

Seidenberg, MacDonald, & Saffran 2002

Our understanding of the contribution of statistical learning is limited by incomplete knowledge of the kinds of statistics infants encode and whether these are the ones relevant to natural language.

Studies centered on understanding rule learning raise two major unresolved issues. First, the distinction between a rule and a statistical generalization remains unclear. Many of the regularities summarized in the table could be called either rules or statistics. Second, how do infants actually find rules in the speech they hear? Is there a procedure that would yield the right rules under realistic learning conditions? The evidence for rule learning is mostly negative: cases where learning occurs but there is no obvious statistical explanation. A theory explaining how rule learners arrive at exactly the correct generalizations given the complexities of their experience would represent substantial progress.

1. Words have PU_KI, BE_GA, TA_DU structure (Peña et al.'s rule, AXC)
2. Initial syllables begin with a stop consonant
3. Final syllables begin with a stop consonant
4. Continuant consonants occur word medially
5. PU predicts RA
6. PU is not followed by BE
7. KI predicts BE
8. PURA predicts KI
9. RAKI predicts BE
10. PUBE is not followed by KI

Statistical learning offers another potential explanation insofar as languages may exhibit only those structures that learners are able to track. Thus, the structure of language may have resulted in part from constraints imposed by the limits of human learning.

Marcus & Berent 2003: On Seidenberg et al. 2002

Arguing against a "reconcilist" position in which complex cognitive functions would depend on a mixture of statistical and algebraic (rule) mechanisms (*1*, *2*), Seidenberg *et al.* favor a position that they describe as "statistical learning," wherein languages are a product not of language-specific knowledge, but of limits on the statistical structures that "learners are able to track."

Unfortunately, nowhere do they spell out what exactly statistical learning consists of. Broadening the notion of statistics from things like transitional probabilities between particular elements (*3*) to relationships between any kind of information, concrete or abstract, trivializes the very term, rendering it broad enough to encompass any lawful relationship, including the very rules that Seidenberg and his colleagues have argued against (*4*). Without a notion of what would not count as statistical learning, it is hard to even see what the hypothesis is;

Was this a trivialization of statistical learning?

What would be an alternative hypothesis?

Marcus & Berent 2003: On Seidenberg et al. 2002

One such hypothesis is that learners might be able to extract and generalize rules, where rules are defined as operations over variables. For example, a simple rule of reduplication might state that **X** goes to **XX**, where **X** is a variable that can stand for a large class of elements (e.g., *b*, *d*, *f*).

Because such rules make reference to variables (e.g., **X**), it follows that speakers should be able to generalize them across the board, to any representable element that can be substituted into the variable, irrespective of the properties of specific elements, their similarity to trained items, and their previous history of statistical cooccurrence (*6*, *7*).

Empirical data suggest that people can indeed generalize in just this way. In addition to being able to learn to recognize statistical relations between particular sets of elements, listeners can also acquire formal patterns that hold for any element, irrespective of its statistical properties, just as the "rule" theory predicts.

Hebrew speakers, for example, recognize that root morphemes that follow an **XYY** (e.g., *sll*, *bdd*) pattern are well formed, whereas roots that follow an **XXY** pattern (e.g., *ssl*, *bbd*) are not, and they extend this generalization to novel word forms (*8*), even for those that contain phonetic contrasts that do not appear in Hebrew (*9*). Similarly, human infants that

Similarly, human infants that have been exposed to sentences like *la ta la* and *ga na ga* appear to recognize the differences between novel items like *wo fe wo* (which follows the same pattern) and *wo fe fe* (which does not) (*10*).

Idea: rules over variables aren't learnable with statistics?

Marcus & Berent 2003: On Seidenberg et al. 2002

Such generalizations are naturally handled by computational systems that come equipped with operations over variables but cannot be captured by systems that are only capable of counting transitional probabilities between known elements, nor, we suspect, by any system that could be reasonably construed as purely statistical (*11*), unless the notion of "statistical" were broadened to the point of being unfalsifiable.

Or at least not by transitional probabilities alone…are there other options that could work in these cases?

But there is no evidence that humans can learn particularly complex statistics (*12*) or that they are uniquely gifted statistical learners—cotton-top tamarins, for example, are just as capable as humans in learning transitional probabilities (*13*). There is little doubt that people can detect correlations and transitional probabilities, but such tools are unlikely to be the only elements in the cognitive equation.

…keeping in mind the limits of human statistical abilities?

Seidenberg, MacDonald, & Saffran 2003: On Marcus & Berent 2003

THE MARCUS AND BERENT LETTER IS A misreading of our Perspective, which was not an argument against a "reconcilist" position or for an exclusively statistical approach to language learning. This is evident from the title, which posed a question rather than making an assertion, and from the text, which pointed out unknowns concerning both grammar-based and statistically based approaches.

Marcus and Berent complain that there is no definition of statistical learning in our Perspective, but our point was that there are ambiguities on both sides that made it difficult to sustain M. Peña *et al.*'s claim to have discovered evidence for two distinct mechanisms, one statistical and one rule-based

In each case, evidence is provided that observed behavior cannot be explained by a specific statistical analysis; it is then inferred that no statistical analysis is viable and that a rule-learning mechanism must therefore exist. Of course, these conclusions are valid only if the behavior does not afford other statistical analyses.

important

...but then, what might those be?

With respect to the evidence cited by Marcus and Berent, at issue is their assumption that if the generalization that humans extract from the data is abstract (e.g., referring to word position rather than adjacent elements), then the behavior reflects rule learning rather than statistics.

Summary: Seidenberg et al. vs. Marcus & Berent

Cognition is more than simple statistical learning like transitional probabilities and correlation.

Cognition requires the ability to abstract rules.

It isn't clear what level of information humans can statistically track.

Seidenberg & friends: Statistical learning will be able to track the things traditionally perceived as rules. That is, statistical information will be over abstract relations as well as basic level things like words and syllables.

Marcus & Berent: No, it won't. Rule learning is about symbolic manipulation. And if you say statistical learning can be over abstract symbolic units, then it's not just simple statistical learning.

…and the jury is still out.

Discussion Questions

Statistical learning vs. rules: what's the difference? Can these be combined? How would/could these interact?

Does statistical learning ever stop, per se? Is it always going on, even if rules are being learned? Is it equivalent to simply finding the dominant pattern, however that pattern is defined?

How would statistical learning work in cases where children are learning more than one language? What needs to happen for learning to be successful in this situation?