Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies

Luca Onnis (lo35@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY 14853, USA

Padraic Monaghan (P.Monaghan@psych.york.ac.uk)
Department of Psychology, University of York, York, YO10 5DD, UK

Morten H. Christiansen (mhc27@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY 14853, USA

Nick Chater (nick.chater@warwick.ac.uk)
Institute for Applied Cognitive Science and Department of Psychology, University of Warwick, Coventry, CV47AL, UK

Abstract

An important aspect of language acquisition involves learning the syntactic nonadjacent dependencies that hold between words in sentences, such as subject/verb agreement or tense marking in English. Despite successes in statistical learning of adjacent dependencies, the evidence is not conclusive for learning nonadjacent items. We provide evidence that discovering nonadjacent dependencies is possible through statistical learning, provided it is modulated by the variability of the intervening material between items. We show that generalization to novel syntactic-like categories embedded in nonadjacent dependencies occurs with either zero or large variability. In addition, it can be supported even in more complex learning tasks such as continuous speech, despite earlier failures.

Introduction

Statistical learning – the discovery of structural dependencies through the probabilistic relationships inherent in the raw input – has long been proposed as a potentially important mechanism in language development (e.g. Harris, 1955). Efforts to employ associative mechanisms for language learning withered during following decades in the face of theoretical arguments suggesting that the highly abstract structures of language could not be learned from surface level statistical relationships (Chomsky, 1957). Recently, interest in statistical learning as a contributor to language development has reappeared as researchers have begun to investigate how infants might identify aspects of linguistic units such as words, and to label them with the correct linguistic abstract category such as VERB. Much of this research has focused on tracking dependencies between *adjacent* elements. However, certain key relationships between words and constituents are conveyed in nonadjacent (or remotely connected) structure. In English, linguistic material may intervene between auxiliaries and inflectional morphemes (e.g., *is cooking*, *has traveled*) or between subject nouns and verbs in number agreement (*the books on the shelf are dusty*). The presence of embedding and nonadjacent relationships in language was a point of serious difficulty for early associationist approaches. It is easy to see that a distributional mechanism computing solely neighbouring information would parse the above sentence as …**the shelf is dusty*. Despite the importance of detecting remote dependencies, we know relatively little about the conditions under which this skill may be acquired by statistical means.

In this paper, we present results using the Artificial Language Learning (ALL) paradigm designed to test learning of nonadjacent dependencies in adult participants. We suggest that a single statistical mechanism might underpin two language learning abilities: detection of nonadjacencies and abstraction of syntactic-like categories from nonadjacent distributional information.

Despite the fact that both infants and adults are able to track transitional probabilities among adjacent syllables (Saffran, Aslin, & Newport, 1996), tracking nonadjacent probabilities, at least in uncued streams of syllables, has proven elusive in a number of experiments and the evidence is not conclusive (Newport & Aslin, 2004; Onnis, Monaghan, Chater, & Richmond, submitted; Peña, Bonatti, Nespor, & Mehler, 2002). Thus, a serious empirical challenge for statistical accounts of language learning is to show that a distributional learner can learn dependencies at a distance. Previous work using artificial languages (Gómez, 2002) has shown that the variability of the material intervening between dependent elements plays a central role in determining how easy it is to detect a particular dependency. Learning improves as the variability of elements that occur between two dependent items increases. When the set of items that participate in the dependency is small relative to the set of elements intervening, the nonadjacent dependencies stand out as invariant structure against the changing background of more varied material. This effect also holds when there is no variability of intervening material shared by different nonadjacent items, perhaps because the intervening material becomes invariant with respect to the variable dependencies (Onnis,

Christiansen, Chater, & Gómez, 2003). In natural language, different structural long-distance relationships such as singular and plural agreement between noun and verb may in fact be separated by the same material (e.g. *the books on the shelf are dusty* versus *the book on the shelf is dusty*). We call the combined effects of zero and large variability the *variability hypothesis*.

Very similar ALL experiments tested have failed to show generalization from statistical information unless additional perceptual cues such as pauses between words were inserted, suggesting that a distributional mechanism alone is too weak to support abstraction of syntactic-like categories. On these grounds Peña et al. (2002) have argued that generalization necessitates a rule-based computational mechanism, whereas speech segmentation relies on lower-level statistical computations. However, these experiments tested nonadjacency learning and embedding generalization with low variability of embedded items, which we contend is consistent with the variability hypothesis that learning should be hard. Our aim is to show that at the end-points of the variability continuum, i.e. with either no or large variability, generalization becomes possible. In Experiment 1, we present results suggesting that both detection of nonadjacent frames and generalization to the embedded items are simultaneously achieved when either one or a large number of different type items are shared by a small number of highly frequent and invariant frames. In Experiment 2 we also investigate whether tracking nonadjacent dependencies can assist speech segmentation and generalization simultaneously, given the documented bias for segmenting speech at points of lowest transitional probability (Saffran et al. 1996a,b).

We conclude that adult learners are able to track both adjacent and nonadjacent structure, and the success is modulated by variability. This is consistent with the hypothesis that a learning mechanism uses statistical information by capitalizing on stable structure for both pattern detection and generalization (Gómez, 2002, Gibson, 1991).

Generalising under variability

The words of natural languages are organized into categories such as ARTICLE, PREPOSITION, NOUN, VERB, etc., that form the building blocks for constructing sentences. Hence, a fundamental part of a language knowledge is the ability to identify the category to which a specific word, say *apple*, belongs and the syntactic relationships it holds with adjacent as well as nonadjacent words. Two properties of word class distribution appear relevant for a statistical learner. First, closed class words like articles and prepositions typically involve highly frequent items belonging to a relatively small set (*am*, *the*, *-ing*, *-s*, *are*) whereas open class words contain items belonging to a very large set (e.g. nouns, verbs, adjectives). Secondly, Gómez (2002) noted that sequences in natural languages involve members of the two broad categories being interspersed. Crucially, this asymmetry translates into patterns of highly invariant *nonadjacent* items, or frames, separated by highly variable material (*am cooking*, *am working*, *am going*, etc.). Such sequential asymmetrical properties of natural language may help learners solve two complex tasks: a) building syntactic constructions that sequentially span one or several words; b) building relevant abstract syntactic categories for a broad range of words in the lexicon that are distributionally embedded in such nonadjacent relationships. Frequent nonadjacent dependencies are fundamental to the process of progressively building syntactic knowledge of, for instance, tense marking, singular and plural markings, etc. For instance, Childers & Tomasello (2001) tested the ability of 2-year-old children to produce a verb-general transitive utterance with a nonce verb. They found that children were best at generalizing if they had been mainly trained on the consistent pronoun frame *He`s VERB-ing it* (e.g., *He`s kicking it*, *He`s eating it*) rather than on several utterances containing unsystematic correlations between the agent and the patient slots (*Mary`s kicking the ball*, *John`s pushing the chair*, etc.).

Gómez (2002) found that the structure of sentences of the form $A_iX_jB_i$, where there were three different $A_i\_B_i$ pairs, could in fact be learned provided there was sufficient variability of X_j words. The structure was learned when 24 different Xs were presented, but participants failed to learn when Xs varied from sets of 2, 4, 6, or 12, i.e. with low variability. Onnis et al. (2003) replicated this finding and also found that learning occurred with only one X being shared, suggesting the nonadjacent structure would stand out again, this time as variant against the invariant X.

While Gómez interpreted her results as a learning bias towards what changes versus what stays invariant, thus leading to "discard" the common embeddings in some way, we argue here that there may be a reversal effect in noting that common elements all share the same contextual frames. If several words – whose syntactic properties and category assignment are *a priori* unknown – are shared by a number of contexts, then they will be more likely to be grouped under the same syntactic label, e.g. VERB. For instance, consider a child faced with discovering the class of words such as *break*, *drink*, *build*. As the words share the same contexts below, s/he may be driven to start extracting a representation of the VERB class (Mintz, 2002):

I am-X-ing
dont-X-it
Lets-X-now!

Mintz (2002) argued that most importantly, in hearing a new word in the same familiar contexts, for instance *eat* in *am-eat-ing*, the learner may be drawn to infer that the new word is a VERB. Ultimately, having categorized in such a way, the learner may extend the usage of *eat* as a VERB to new syntactic constructions in which instances of the category VERB typically occur. For instance s/he may produce a novel sentence *Lets-eat-now!* Applying a category label to an word (e.g. *eat* belongs to VERB) greatly enhances the generative power of the linguist system, because the labeled item can now be used in new syntactic contexts where the category applies. In Experiment 1 we tested whether

generalization to new X items in the *A_X_B* artificial grammar used by Gómez (2002) and Onnis et al. (2003) is supported under the same conditions of no or large variability that affords the detection of invariant structure. Hence, if frames are acquired under the variability hypothesis, generalization will be supported when there is either zero or large variability of embeddings. Likewise, because invariant structure detection is poor in conditions of middle variability, generalization is expected to be equally poor in those conditions too.

Experiment 1

Method

Subjects

Thirty-six undergraduate and postgraduate students at the University of Warwick participated and were paid £3 each.

Materials

In the training phase participants listened to auditory strings generated by one of two artificial languages (L1 or L2) of the type $A_iX_jB_i$. Strings in L1 had the form $A_1X_jB_1$, $A_2X_jB_2$, and $A_3X_jB_3$. L2 strings had the form $A_1X_jB_2$, $A_2X_jB_3$, $A_3X_jB_1$. Variability was manipulated in 3 conditions – zero, small, and large– by drawing X from a pool of either 1, 2 or 24 elements. The strings, recorded from a female voice, were the same that Gómez used in her study and were originally chosen as tokens among several recorded sample strings in order to eliminate talker-induced differences in individual strings.

The elements A_1, A_2, and A_3 were instantiated as *pel*, *vot*, and *dak*; B_1, B_2, and B_3, were instantiated as *rud*, *jic*, *tood*. The 24 middle items were *wadim*, *kicey*, *puser*, *fengle*, *coomo*, *loga*, *gople*, *taspu*, *hiftam*, *deecha*, *vamey*, *skiger*, *benez*, *gensim*, *feenam*, *laeljeen*, *chla*, *roosa*, *plizet*, *balip*, *malsig*, *suleb*, *nilbo*, and *wiffle*. The middle items were stressed on the first syllable. Words were separated by 250-ms pauses and strings by 750-ms pauses. Three strings in each language were common to all two groups and they were used as test stimuli. The three L2 items served as foils for the L1 condition and vice versa. The test stimuli consisted of 12 strings randomized: six strings were grammatical and six were ungrammatical. The ungrammatical strings were constructed by breaking the correct nonadjacent dependencies and associating a head to an incorrectly associated tail, i.e. $*A_iXB_j$.. Six strings (three grammatical and three ungrammatical) contained a previously heard embedding, while 6 strings (again three grammatical and three ungrammatical) contained a new, unheard embedding. Note that correct identification could only be achieved by looking at nonadjacent dependencies, as adjacent transitional probabilities were the same for grammatical and ungrammatical items.

Procedure

Six participants were recruited in each of 3 Variability conditions (1, 2 and 24) and for each of two Language conditions (L1, L2) resulting in 12 participants per Variability condition. Learners were asked to listen and pay close attention to sentences of an invented language and they were told that there would be a series of simple questions relating to the sentences after the listening phase. During training, participants in the two conditions listened to the same overall number of strings, a total of 432 token strings. This way, frequency of exposure to the nonadjacent dependencies was held constant across conditions. Participants in set-size 24 heard six iterations of each of 72 type strings (3 dependencies x 24 middle items), participants, in set-size 2 encountered each string 12 times as often as those exposed to set size 24, and so forth. Hence, whereas nonadjacent dependencies where held constant, transitional probabilities of adjacent items decreased as set size increased.

Training lasted about 18 minutes. Before the test, participants were told that the sentences they had heard were generated according to a set of rules involving word order, and they would now hear 12 strings, 6 of which would violate the rules. They were asked to give a "Yes/No" answer. They were also told that the strings they were going to hear may contain new words and they should base their judgment on whether the sentence was grammatical or not on the basis of their knowledge of the grammar. This is to guarantee that participants did not select as ungrammatical all the sentences with novel words simply because they contained novel words.

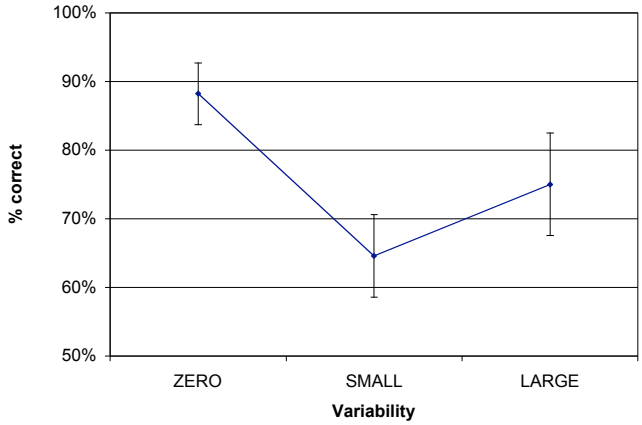

Figure 1. Generalisation under variability - Exp.1

Results and discussion

An analysis of variance with Variability (1 vs. 2 vs. 24) and Language (L1 vs. L2) as between-subjects and Grammaticality (Trained vs. Untrained strings) as a within-subjects variable resulted in a main Variability effect, $F(2,30)= 3.41$, $p< .05$, and no other interaction. Performance across the different variability conditions resulted in a U-shaped function: a polynomial trend analysis showed a significant quadratic effect, $F(1, 35) =7.407$, $p <.01$. Figure 1 presents the percentage of endorsements for total accuracy in each of the three variability conditions. These results add considerable power to the variability hypothesis: not only can nonadjacencies be detected, but generalization too can occur distributionally, and *both* processes seem to be modulated by the same conditions of variability. In addition, generalization with zero variability allows us to disambiguate previous results, in that the high performance obtained by Onnis et al. (2003) could have been due to a simple memorization of the 3 strings repeated over and over

again during training. However, in Experiment 1 correct classification of new strings as grammatical can only be done on the basis of the correct nonadjacencies. Thus, it seems that learning on zero or large variability conditions is supported by a similar mechanism. Finally, we note that A and B words are monosyllabic and X words are bysillabic, participants could simply learn a pattern S-SS-S (where S=syllable). However, because all sentences display such pattern across conditions this cannot explain the U-shape of the learning curve.

Experiment 2

In Experiment 1 the items of the grammar are clearly demarcated by pauses. It can be argued that this makes the task somewhat simplified with respect to real spoken language, which does not contain for instance such apparent cues at every word boundary. In addition, the embedded item *X* was instantiated in bisyllabic words (as opposed to monosyllabic *A* and *B* words), providing an extra cue for category abstraction. In this context, Peña et al. (2002) have argued that generalization and speech segmentation are separate processes underpinned by separate computational mechanisms: statistical computations are used in a segmentation task but this is not performed simultaneously with algebraic computations that would permit generalizations of the structure. Once the segmentation task was solved by introducing small pauses in the speech signal, their underlying structure was learned. Hence it is important to test these claims in the light of the variability hypothesis, which we argue might provide the key to learning nonadjacencies and generalizing altogether, even in connected speech, without invoking two separate mechanisms.

Recent attempts to show statistical computations of a higher order at work in connected speech with a similar AXB language have met with some difficulty: Newport & Aslin (2004), for instance, exposed adults to a continuous speech stream, created by randomly concatenating *AXB* words with 3 *A_B* syllable dependencies and with 2 different middle *X* syllables. A sample of the speech stream obtained would be …$A_1X_3B_1A_2X_2B_2A_3X_1B_3$…. In this case participants were unable to learn the nonadjacent dependencies. Concatenating words seamlessly adds considerable complexity to the task of tracking statistical information in the input for two main reasons: first, transitional probabilities between words of a language containing, say 3 dependencies and 3 Xs, p($B|A$)= 0.5 are higher than within words, p($X|A$) and p($B|X$)= 0.33, and this pressures for segmentation within words (Saffran, Aslin, & Newport, 1996ab). Secondly, assuming the statistical mechanism is sensitive to nonadjacent dependencies as seems the case in Experiment 1, concatenating items entails the additional burden of tracking nonadjacent transitional probabilities across word boundaries, e.g. $X_3\_A_2$, $B_1\_X_2$, and dependencies spanning *n* words away can in principle also be attended to, e.g. two items away ($B_2\_\ \_A_3$…,etc.). One can readily see that if all transitional probabilities of different order were to be computed this scenario would soon create a computational impasse. The insight from Gómez (2002) and Experiment 1 is that variability plays a key role, in that it allows adjacent dependencies to be overcome in favour of nonadjacent ones, but it remains to be seen whether this can be done in connected speech too.

Peña et al. (2002) tested participants on whether they learned to generalize from the rules of an *AXB* language very similar to Newport & Aslin (2004) in unsegmented speech. Again *AXB* items were instantiated in syllables and formed words concatenated one to the other seamlessly. At test, participants demonstrated no preference for so-called "rule-words", new trigram sequences that maintained the $A_i\_B_i$ nonadjacent dependencies but contained a different *A* or *B* in the intervening position (e.g., $A_1B_3B_1$), compared to part-words, i.e., sequences that spanned word boundaries (e.g., $X_2B_1A_3$, or $B_3A_1X_2$). In a further manipulation, 25-ms gaps were introduced between words during the training phase of the experiment, and now participants generalized as indicated by a preference for rule-words over part-words. Peña et al. claimed that altering the speech signal resulted in a change in the computations performed by their participants. Statistical computations were used in a (previously successful) segmentation task but this was not performed simultaneously with algebraic computations that would permit generalizations of the structure. They argued that once the segmentation task was solved by introducing small gaps in the speech signal, the underlying structure would be learned. However, using the same stimuli and experimental conditions as Peña et al. Onnis, Monaghan, Chater & Richmond (submitted) found that rule-words were preferred over part-words in both segmentation and generalization tasks even when the nonadjacent structure was eliminated: participants reliably preferred incorrect rule-words $*A_1B_3B_2$ to part-words B_1A_2X, due to preference for plosive sounds in word-initial position. Hence such preference did not reflect learning of nonadjacent dependencies. Although discouraging at first sight, all these negative results are not inconsistent with the variability hypothesis. In fact, they are all cases structurally similar to the low-variability condition in Gómez (2002) and Experiment 1. Thus, in Experiment 2 we tested whether with sufficiently large variability:

a) tracking higher-order dependencies can be used to segment speech. This is a difficult task because it implies overriding even lower transitional probabilities p($X|A$) than previously tested and this pressures for segmentation within word boundaries (Saffran et al. 1996);

b) generalization of the embeddings can occur *simultaneously* to speech segmentation, i.e. on-line in running speech, and can be done by statistical analysis of the input alone, i.e. without additional perceptual cues such as pauses. We tested this using the same material and training conditions as Peña et al. for their unsuccessful pause-free generalization task, but increasing the variability of the *X* syllables to 24 items as in Experiment 1.

Method

Subjects

20 undergraduate and postgraduate students at the University of Warwick participated for £1. All participants spoke English as a first language and had normal hearing.

Materials

We used the same nine word types from Peña et al.'s Experiment 2 to construct the training speech stream in our Experiment 2. The set of nine words was composed of three groups ($A_i\_B_i$), where the first and the third syllable were paired, with an intervening syllable (X) selected from one of either three syllables (low variability condition) or 24 syllables (high variability condition). The syllables were randomly generated from the following set of consonants: /p/,/b/,/g/,/k/,/d/,/t/,/l/,/r/,/f/,/tʃ/,/dʒ/,/n/, /s/,/v/,w/,/m/,/θ/,/ʃ/,/z/and the following vowels: /ɛi/,/uw/,/a/,/iy/,/au/,/oi/,/ai/,/æ/,/œ/.

Consonants and vowels were permuted, then joined together. No syllables occurred more than once in the set of 33 generated. Each participant listened to a different permutation of consonant-vowel pairings. Notice that the language structure in the two conditions match very closely those of small and large variability in Experiment 1. Unlike Experiment 1 all items were monosyllabic and equally stressed.

Words were produced in a seamless speech stream, with no two words from the same set occurring adjacently, and no same middle item occurring in adjacent words. Hence, adjacent transitional probabilities were as follows: for the small variability condition, and within words, $p(\underline{X}|\underline{A})$ and $p(\underline{B}|\underline{X})= 0.33$; between adjacent words $p(\underline{B_i}|\underline{A_i})= 0.5$. Nonadjacent transitional probabilities were $p(B_i|A_i)= 1$, $p(A_i|X_{previous})= 0.33$, $p(X_j|B_{previous})= 0.33$. For the large variability condition all probabilities were the same except within word adjacent probabilities $p(X|A) = 0.041$.

Therefore, the predicition is that if learners computed adjacent statistical probabilities they should prefer part-words and perhaps significantly more in the large variability condition. Conversely, if they computed nonadjacent dependencies they would rely on the most statistically reliable ones, namely $p(B_i|A_i)= 1$, i.e. they would segment correctly at word boundary.

We used the Festival speech synthesizer using a voice based on British-English diphones at a pitch of 120 Hz, to generate a continuous speech stream lasting approximately 10 minutes. All syllables were of equal duration, and were produced at a rate of 4.5 syllables/second. Words were selected randomly, except that no $A_i\_B_i$ pair occurred twice in succession. The speech stream was constructed from 900 words, in which each word occurred approximately 100 times. The speech stream faded in for the first 5 seconds, and faded out for the last 5 seconds, so there was no abrupt start or end to the stream. In addition, and crucially, for each participant, we randomly assigned the 9 syllables from the first experiment to the A_i, B_i and X_j positions. Thus, each participant listened to speech with the same structure containing the nonadjacent dependencies, but with syllables assigned to different positions. This was to avoid any bias towards choosing a rule-word because of a preference for plosive sounds, as Onnis et al. (submitted) demonstrated. Part-words were formed from the last syllable of one word and two syllables from the following word (B_iA_jX), or from the last two syllables of one word and the first syllable from the following word (XB_iA_j).

Procedure

In the training phase, participants were instructed to listen to continuous speech and try and work out the "words" that it contained. They then listened to the training speech. At test part-words were compared to "rule-words", which were composed of $A_i\_\_B_i$ pairs with an intervening item that was either an A_j or a B_j from another $A_j\_B_j$ pair. Participants were requested to respond which of two sounds was a "word" in the language they had listened to. They were then played a "rule-word" and a part-word separated by 500 ms, and responded by pressing either "1" on a computer keyboard for the first sound a word, or "2" for the second sound a word. After 2 seconds, the next rule-word and part-word pair were played. In half of the test trials, the "rule-words" occurred first. Five participants heard a set of test trials with one set of words first, and the other 5 participants heard the other set of words first.

Results

The results are shown in Figure 2. In line with the original Peña et al.'s experiment, we found no evidence for participants learning to generalize from the nonadjacent structure of the stimuli in the low-variability condition. Participants responded with a preference for rule-words over part-words 41.9% of the times, which was significantly lower than chance, $t(9) = -2.73$, $p < .05$. Conversely, in the high-variability condition participants preferred rule-words 63.3% of the times, significantly higher than chance, $t(9)= -3.80$, $p = .0042$. In addition, there was a significant difference between the low variability and the high variability condition, $t(18) = -4.68$, $p < .001$.

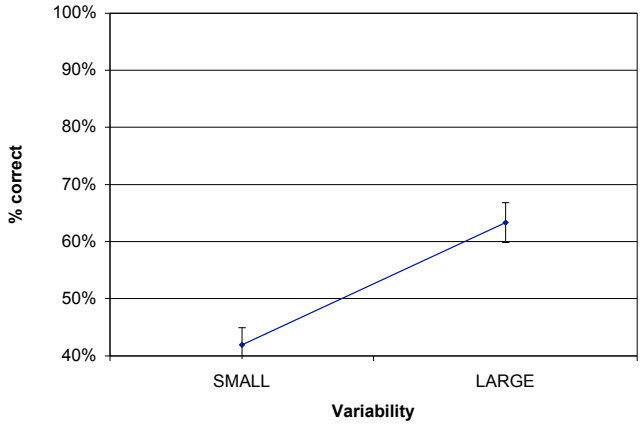

Figure 2. Generalisation in unsegmented speech - Exp. 2

General Discussion

Statistical learning of dependencies between adjacent elements in a sequence is fast, robust, automatic and general in nature. In contrast, although the ability to track remote dependencies is a crucial linguistic ability, relatively little research has been directed toward this problem. Nonadjacent structure in sequential information seems harder to learn, possibly because learners have to overcome the bias toward adjacent transitional probabilities. In fact, a statistical learning mechanism that kept track of all possible adjacent and nonadjacent regularities in the input, including syllables one, two, three away, etc., would quickly

encounter a computationally intractable problem of exponential growth. It would seem that either statistical learning is limited to sensitivity to adjacent items, or there may be statistical conditions in which adjacencies become less relevant in favour of nonadjacencies. It has been suggested that this applies under conditions of large variability of the intervening material (Gómez, 2002) or zero variability (Onnis et al., 2003). This paper contributes some steps forward: first, Experiment 1 shows that variability is the key not only for detection of remote dependencies but also for generalization of embedded material, fostering the creation of abstract syntactic-like classes, which is often assumed to require higher-level algebraic computation. Secondly, in Experiment 2 segmentation and generalization are achieved simultaneously, without the assist of pauses (a difference in signal) as Pena et al. claimed. Consequently, rather than supporting a statistical/algebraic distinction our results suggest specific selectivities in learning patterned sequences. The specific characterization of such selectivities may not be simple to identify: Newport & Aslin (2004) found that nonadjacent segments (consonants and vowels) could be learned but not nonadjacent syllables, and proposed that this accounts for why natural languages display nonadjacent regularities of the former kind but not of the latter. Experiment 2, however, shows that with large variability nonadjacent syllabic patterns can in fact be learned. The key factor for success is again variability. Experiment 2 also shows that learners are indeed able to track nonadjacent dependencies in running speech, despite the well documented bias for adjacent associations and the preference for segmenting continuous speech at points of lowest transitional probabilities.

Overall, the results suggest that the learning mechanism entertains several statistical computations and implicitly "tunes in" to statistical relations that yield the most reliable source of information. This hypothesis was initiated by Gómez (2002) and is consistent with several theoretical formulations such as reduction of uncertainty (Gibson, 1991) and the simplicity principle (Chater, 1996) that the cognitive system attempts to seek the simplest hypothesis about the data available. In the face of performance constraints and way too many statistical computations, the cognitive system may be biased to focus on data that will be likely to reduce uncertainty. Specifically, whether the system focuses on transitional probabilities or nonadjacent dependencies may depend on the statistical properties of the environment that is being sampled.

Our work ties in with recent acquisition literature that has emphasized the constructive role of syntactic frames as the first step for building more abstract syntactic representations (Tomasello, 2003 for an overview). Children's syntactic development would build upon several consecutive stages from holophrases such as *I-wanna-see-it* (at around 12 months), to pivot-schemas (*throw-ball*, *throw-can*, *throw-pillow*, at about 18 months), through item-based constructions (*John hugs Mary*, *Mary hugs John*, at about 24 months), to full abstract syntactic constructions (*a X*, *the Xs*, *Eat a X*).

Statistical learning seems, at least in adults, powerful enough to allow the discovery of complex nonadjacent structure, but simply not any condition will do: we have suggested that variability such as that emerging from the asymmetry between open and closed class words may be a crucial ingredient for understanding the building of language.

Acknowledgments

We thank M. Merkx for running Exp. 2, and R. Gómez for the stimuli in Exp.1 and important insights. Part of this work was conducted while L. Onnis and P. Monaghan were at the University of Warwick. Support comes from European Union Project HPRN-CT-1999-00065, and Human Frontiers Science Program.

References

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, 103, 566-581.

Childers, J. & Tomasello, M. (2001). The role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology*, *37*, 739-748.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Gibson, E.J. (1991). *An Odyssey in Learning and Perception.* Cambridge, MA: MIT Press.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431-436.

Harris, Z.S. (1955). From phoneme to morpheme. *Language 31*, 190-222.

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

Newport, E.L., & Aslin, R.N. (2004). Learning at a distance I. Statistical learning of nonadjacent dependencies. Cognitive Psychology, *48*, 127-162.

Onnis, L., Monaghan, P., Chater, N., & Richmond, K. (submitted). Phonology impacts segmentation and generalization in speech processing.

Onnis, L., Christiansen, M., Chater, N., & Gómez, R. (2003). Reduction of uncertainty in human sequential learning: Evidence from Artificial Grammar Learning. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, 887-891.

Peña, M., Bonatti, L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298,* 604-607.

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

Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606-621.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.