Reduction of Uncertainty in Human Sequential Learning: Evidence from Artificial Grammar Learning

Luca Onnis (l.onnis@warwick.ac.uk)
Department of Psychology, University of Warwick, Coventry, CV47AL, 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

Rebecca Gómez (rgomez@u.arizona.edu)
Department of Psychology, University of Arizona, Tucson, AZ 85721, USA

Abstract

Research on statistical learning in adults and infants has shown that humans are particularly sensitive to statistical properties of the input. Early experiments in artificial grammar learning, for instance, show a sensitivity for transitional n-gram probabilities. It has been argued, however, that this source of information may not help in detecting nonadjacent dependencies, in the presence of substantial variability of the intervening material, thus suggesting a different focus of attention involving change versus non-change (Gómez, 2002). Following Gómez proposal, we contend that alternative sources of information may be attended to simultaneously by learners, in an attempt to reduce uncertainty. With several potential cues in competition, performance crucially depends on which cue is strong enough to be relied upon. By carefully manipulating the statistical environment it is possible to weigh the contribution of each cue. Several implications for the field of statistical learning and language development are drawn.

Introduction

Research in artificial grammar learning (AGL) and artificial language learning (ALL) in infants and adults has revealed that humans are extremely sensitive to the statistical properties of the environment they are exposed to. This has opened up a new trend of investigations aimed at determining empirically the processes involved in so-called statistical learning.

Several mechanisms have been proposed as the default that learners use to detect structure, although crucially there is no consensus in the literature over which is most plausible or whether there is a default at all. Some researchers have shown that learners are particularly sensitive to transitional probabilities of bigrams (Saffran, Aslin, & Newport, 1996): confronted with a stream of unfamiliar concatenated speech-like sound they tend to infer word boundaries between two syllables that rarely occur adjacently in the sequence. Sensitivity to transitional probabilities seems to be present across modalities, for instance in the segmentation of streams of tones (Saffran, Johnson, Aslin, and Newport, 1999) and in the temporal presentation of visual shapes (Fiser & Aslin, 2002).

Other researchers have proposed exemplar- or fragment-based models, based on knowledge of memorised chunks of bigrams and trigrams (Dulany et al., 1984; Perruchet & Pacteau, 1990; Servan-Schreiber & Anderson, 1990) and learning of whole items (Vokey & Brooks, 1992). Yet others have postulated rule-learning in transfer tasks (Reber, 1967; Marcus, Vijayan, Rao & Voshton, 1999). In addition, knowledge of chained events such as sentences in natural languages require learners to track nonadjacent dependencies where transitional probabilities are of little help (Gómez, 2002).

In this paper we propose that there may be no default process in human sequential learning. Instead, learners may be actively engaged in search for good sources of reduction in uncertainty. In their quest, they seek alternative sources of predictability by capitalizing on information that is likely to be the most statistically reliable. 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. Given performance constraints, the cognitive system may be biased to focus on data that will be likely to reduce uncertainty as far as possible[1]. Specifically, whether the system focuses on transitional probabilities or non-adjacent dependencies may depend on the statistical properties of the

[1] We assume that this process of selection is not necessarily conscious, and might for example involve distribution of processing activity in a neural network.

environment that is being sampled. Therefore, by manipulating the statistical structure of that environment, it is perhaps possible to investigate whether active search is at work in detecting structure.

In two experiments, we investigated participants' degree of success at detecting invariant structure in an AGL task in 5 conditions where the test items and test task are the same but the probabilistic environment is manipulated so as to change the statistical landscape substantially. We propose that a small number of alternative statistical cues might be available to learners. We aim to show that, counter to intuition, orthogonal sources of reliability might be at work in different experimental conditions leading to successful or unsuccessful learning. We also asked whether our results are robust across perceptual modalities by running two variations of the same experiment, one in the auditory modality and one in the visual modality. Our experiments are an extension of a study by Gómez (2002), which we first introduce.

Detection of invariant structure through context variability

Many sequential patterns in the world involve tracking nonadjacent dependencies. For example, in English auxiliaries and inflectional morphemes (e.g., *am cooking*, *has travelled*) as well as dependencies in number agreement (*the books on the shelf are dusty*) are separated by various intervening linguistic material. One potential source of learning in this case might be embedding of first-order conditionals such as bigrams into higher-order conditionals such as trigrams. That learners attend to n-gram statistics in a chunking fashion is evident in a number of studies (Schvaneveldt & Gómez, 1998; Cohen, Ivry, & Keele, 1990). In the example above chunking involves noting that *am* and *cook* as well as *cook* and *ing* are highly frequent and subsequently noting that *am cooking* is highly frequent too as a trigram. Hence we may safely argue that higher order n-gram statistics represent a useful source of information for detecting nonadjacent dependencies.

However, sequences in natural languages typically involve some items belonging to a relatively small set (functor words and morphemes like *am, the, -ing, -s, are*) interspersed with items belonging to a very large set (e.g. nouns, verbs, adjectives). Crucially, this asymmetry translates into patterns of highly invariant nonadjacent items separated by highly variable material (*am cooking*, *am working*, *am going*, etc.). Gómez (2002) suggested that knowledge of n-gram conditionals cannot be invoked for detecting invariant structure in highly variable contexts because first-order transitional probabilities, P(Y|X), decrease as the set size of Y increases. Similarly, second-order transitional probabilities, P(Z|XY), also decrease as a function of set size of X. Hence, statistical estimates for these transitional probabilities tend to be unreliable. Gómez exposed infants and adult participants to sentences of an artificial language of the form *A-X-B*. The language contained three families of nonadjacent pairs, notably A_1—B_1, A_2—B_2, and A_3—B_3. She manipulated the set size of the middle element X in four conditions by systematically increasing the number from 2 to 6 to 12 and 24 word-like elements. In this way, conditional bigram and trigram probabilities decreased as a function of number of middle words. In the test phase, participants were required to subtly discriminate correct nonadjacent dependencies, (e.g. A_2-X_1-B_2) from incorrect ones (*A_2-X_1-B_1). Notice that the incorrect sentences were new as trigrams, although both single words and bigrams had appeared in the training phase in the same positions. Hence the test requires very fine distinctions to be made. Gómez hypothesized that if learners were focusing on n-gram dependencies they should learn nonadjacent dependencies better when exposed to small sets of middle items because transitional probabilities between adjacent elements are higher for smaller than for larger set sizes. Conversely, if learners spotted the invariant structure better in the larger set size, Gómez hypothesized that increasing variability in the context must have led them to disregard the highly variable middle elements. Her results support the latter hypothesis: learners performed poorly with low variability whereas they were particularly good when the set size of the middle item was largest (24 middle items; see Figure 1).

Testing the zero-variability hypothesis

Gómez proposed that both infant and adult learners are sensitive to change versus non-change, and use their sensitivity to capitalize on stable structure. Learners might opportunistically entertain different strategies in detecting invariant structure, driven by a reduction of uncertainty principle. In this study we are interested in taking this proposal further by exploring what happens when variability between the end-item pairs and the middle items is reversed in the input. Gómez attributed poor results in the middle set sizes to low variability: the *variability effect* seems to be attended to reliably only in the presence of a critical mass of middle items. However, an alternative explanation is that in small set size conditions both nonadjacent dependencies and middle items vary, but none of them considerably more than the other. This may confuse learners, in that it is not clear which structure is non-variant. With larger set sizes middle items are considerably more variable than first-last item pairings, making the nonadjacent pairs stand out as invariant. We asked what happens when variability in middle position is eliminated, thus making the nonadjacent items variable. We replicated Gómez' experiment with adults and added a new condition, namely the zero-variability condition, in which there is only one middle element (e.g. A_3-X_1-B_3, A_1-X_1-B_1). Our prediction is that non-variability of the middle item will

make the end-items stand out, and hence detecting the appropriate nonadjacent relationships will become easier, increasing mean performance rates. Intuitively, sampling transitional probabilities with large context variability results in low information gain as the data are too few to be reliable; by the same vein, the lack of variability should produce low information gain for transitional probabilities as well, because it is just obvious what the bigram structure is. Hence this should make nonadjacent dependencies stand out, as potentially more informative sources of information, by contrast.

The final predicted picture is a U-shape learning curve in detecting nonadjacent dependencies, on the assumption that learning is a flexible and adaptive process.

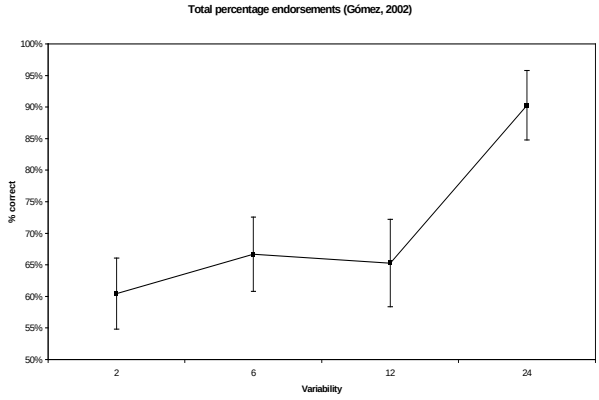

Figure 1. Total percentage endorsements from Gómez (2002) for the different conditions of variability of the middle item.

Experiment 1

Method

Participants Sixty 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). Strings in L1 had the form *aXd*, *bXe*, and *cXf*. L2 strings had the form *aXe*, *bXf*, *cXd*. Variability was manipulated in 5 conditions, by drawing X from a pool of either 1, 2, 6, 12, 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*, *b*, and *c* were instantiated as *pel*, *vot*, and *dak*; *d*, *e*, and *f*, 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*. Following the design by Gómez (2002) the group of 12 middle elements were drawn from the first 12 words in the list, the set of 6 were drawn from the first 6, the set of 2 from the first 2 and the set of 1 from the first word. Three strings in each language were common to all five groups and they were used as test stimuli. The three L2 items served as foils for the L1 condition and vice versa. In Gómez (2002) there were six sentences generated by each language, because the smallest set size had 2 middle items. To keep the number of test items equal to Gómez we presented the 6 test stimuli twice in two blocks, randomizing *within* blocks for each participant. Words were separated by 250-ms pauses and strings by 750-ms pauses.

Procedure Six participants were recruited in each of the five set size conditions (1, 2, 6, 12, 24) and for each of the two language conditions (L1, L2) resulting in 12 participants per set size. 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 all 5 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. For instance participants in set-size 24 heard six iterations of each of 72 type strings (3 dependencies x 24 middle items), participants in set-size 12 encountered each string twice as often as those exposed to set size 24 and so forth. Hence whereas nonadjacent dependencies where held constant, transitional probabilities 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 press "Y" on a keyboard if they thought a sentence followed the rules and to press "N" otherwise.

Results and Discussion

An analysis of variance with Set Size (1 vs. 2 vs. 6 vs. 12 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 effect of Grammaticality, $F(1,50)=24.70$, $p<.001$, a main Set Size effect, $F(4,50)=3.85$, $p<.008$, and a Language x Set Size interaction, $F(4,50)=2.59$, $p<.047$. We were particularly interested in determining whether performance across the different set-size conditions would result in a U-shaped function. Consistent with our prediction, a polynomial trend analysis yielded a significant quadratic effect, $F(1,50)=5.85$, $p<.05$. In

contrast to Gómez (2002), there was not a significant increase between set size 12 and set size 24, t(22)=.57, p=.568. This leveling off is responsible for a significant cubic effect, F(1,50)=9.49, p<.005. Figure 2 summarizes total percentage endorsements for correct answers.

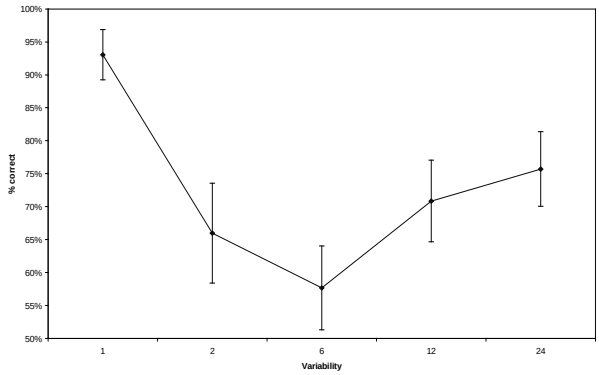

Figure 2. Total percentage endorsements in Experiment 1 for different variability.

Experiment 2

Method

Participants Sixty undergraduate and postgraduate students at the University of Warwick participated and were paid £3 each. None of them had participated in Experiment 1.

Materials. The stimuli were identical to those used in Experiment 1, except that they were presented visually instead of auditorily.

Procedure. Exactly the same procedure as in Experiment 1 was used. Participants sat and looked at the strings as they appeared on the screen. Training lasted approximately 18 minutes, as in Experiment 1. Each string from the language was flashed up in black typeface against white background on a computer screen. Each string stayed on the screen for 2 seconds and was followed by a 750-ms white screen so that the strings could be perceived as independent one from the other. These values were chosen so that training lasted as long as training in Experiment 1. The test phase was the same as in Experiment 1, except that test stimuli were presented visually on the screen.

Results and discussion

An analysis of variance with Set Size (1 vs. 2 vs. 6 vs. 12 vs. 24) and materials (L1 vs. L2) as between-subjects and grammaticality (trained vs. untrained strings) as a within-subjects variable resulted in a main effect of Grammaticality, F(1, 50) =16.39, p <.001, but no significant Grammaticality x Set Size interaction, F(4, 50)=.971, p<.505. There were no other main effects or interactions. In contrast to Experiment 1, a polynomial trend analysis did not show a significant quadratic effect, F<1. Figure 3 presents the percentage of endorsements for total accuracy in each of the five set-size conditions.

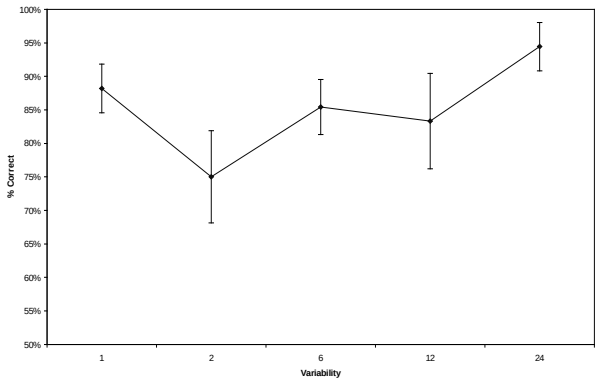

Figure 3. Total percentage endorsements in Experiment 2 for different variability.

General discussion

We used a simple artificial language to enquire into the way learners track remote dependencies. Knowledge of sequence events in the world, including language, involves detecting fixed nonadjacent dependencies interspersed with highly variable material. Gómez (2002) found what we dub a *variability effect*, i.e. a facilitatory effect in detecting invariant structure when the context is highly variable, but not when it is moderately or even little variable. In general, this points to a specific sensitivity to change versus non-change. Conditions 2 to 4 in our Experiment 1 replicate her findings, although performance in terms of percent accuracy seems to improve only gradually from set size 2 to 24, whereas Gómez found a significant difference between set size 12 and 24.

Overall, Gómez' original results do not square well with recent findings of learners' striking sensitivity to n-gram transitional probabilities. Because transitional probabilities are higher in set sizes 2, 6, and 12, performance should be better. Instead, the opposite is the case. We reasoned that perhaps variability in both the middle item and end-point items leave learners in doubt as to what is the invariant structure. Hence, by eliminating variability in the middle item in a new condition, the variability of the nonadjacent items stands out again, this time reversed. However, the effect is, quite counter intuitively, not reversed. Indeed similar performance results are obtained for set size 1 and set size 24. In set size 1 performance is near 100% and significantly better than set size 2 (Experiment 1). One could argue that word trigrams, if recorded perfectly, could suffice to account for performance in set size 1, thus trivializing our results and explaining away the variability effect in this condition. However, as a

counter to that it would be reasonable to expect good performance in set size 2 condition too, given the high number of repetitions (72) for only six type strings. A control condition is currently being run involving learning six frames (instead of three) with 1 different middle item each (e.g. A_3-X_3-B_3, A_6-X_6-B_6) so as to reproduce the same number of type and token frequencies of set size 2, but with no middle item being *shared* by different frames. Similarly, one could argue that good performance in set size 24 could be achieved by strikingly but not impossibly memorizing 72 type strings. However, this would imply good performance in all smaller set sizes as well and this runs counter to data.

Notice also that in all conditions, including set size 1, bigram transitional probabilities by themselves are not sufficient for detecting the correct string *pel wadim rud* from the incorrect one **pel wadim jic* (example taken from L1) as both *pel wadim*, *wadim rud*, and *wadim jic* appear as bigrams during training. Moreover, Gómez (2002) conjectured that perhaps low discrimination rates in small set sizes are due to overexposure of string tokens during training, resulting in boredom and distraction. Our findings disconfirm this hypothesis: if it held true performance would drop even lower in the zero-variability condition, as the type/token ratio decreases even more. Crucially, the finding that there is a statistically significant difference in learning in the two conditions becomes intriguing for several reasons.

A larger project underway examines the extent to which a U-shape learning curve is modality-independent. In Experiment 2 training and test stimuli were presented visually on a computer screen. The obtained U-shape curve is less marked. One possible explanation is that attending to visually presented word-like strings is less demanding cognitively, suggesting a ceiling effect. This explanation is preliminary and needs further evidence. However, the fact that results in Experiment 2 show the same trend as Experiment 1 are encouraging.

The implications of our findings might inform in various degrees both the AGL community and researchers of language development. AGL researchers working mainly with adults have long debated whether there are one or more mechanisms at work in learning structured events from experience. Our results suggest that associative learning based on adjacent material may not be the only source of information. There seems to be a striking tendency to detect variant versus invariant structure, and the way learners do it is extremely adaptive to the informational demands of their input. Without claiming exhaustiveness we explored two putative sources of information, namely n-gram transitional probabilities and the variability effect. At this stage we can only give an informal explanation of the reduction of uncertainty hypothesis. Intuitively, sampling bigrams involving middle items under no variability yields no information gain, as the middle item is always the same. Under this condition learners may be driven to shift attention towards nonadjacent structure. Likewise, sampling bigrams with large variability yields no reduction of uncertainty, as bigram transitional probabilities are very low. In a similar way then, learners may be lead to focus on nonadjacent dependencies. With low variability, sampling bigrams may be reliable enough, hence "distracting" learners away from nonadjacent structure. Other sources may be at work and disentangling the contribution of each of them to learning is an empirical project yet to be investigated. For instance, post-test verbal reports from the majority of our participants suggest that, regardless of their performance, they were aware of the positional dependencies of single words in the strings. This piece of information may be misleading for our task: on the one side it reduces uncertainty by eliminating irrelevant hypotheses about words in multiple positions (each word is either initial, middle, or final), on the other side distinguishing *pel wadim rud* from **pel wadim jic* requires more than positional knowledge. We believe that positional knowledge deserves more research in the current AGL literature. Studies of sequential learning have found that it is an important source of information. However, many nonadjacent dependencies are free ranging and hence non-positionally dependent. Further experiments are needed to investigate whether people can detect such non-positionally dependent constraints as *A_x_y_B*, *A_x_y_w_B*, *A_x_y_w_z_B*, equally well.

Our results have been modeled successfully using a connectionist model. Onnis *et al.* (submitted) use simple recurrent neural networks (SRNs) trained in experimental conditions akin to the adult data reported here, obtaining a very similar U-shape curve. SRNs can be thought of as reducing uncertainty in that predictions tend to converge towards the optimal conditional probabilities of observing a particular successive item to the sequence presented up to that point. The SRNs specific task was to predict the third nonadjacent element B_i correctly. Minimizing the sum squared error maximizes the probability of the next element, given previously occurring adjacent elements (McClelland, 1998). This is equivalent to exploiting bigram probabilities. As we have seen, conditional probability matching only yields suboptimal behaviour. To overcome this, SRNs possess a stack of memory units that help them maintain information about previously encountered material. Crucially, they maintain a trace of the correct non-adjacent item A_i under either no variability or large variability only. This happens by forming separate graded representations in the hidden units for each nonadjacent dependency.

The reduction of uncertainty hypothesis may also be given a formal account in terms of active data selection (MacKay, 1992, Oaksford & Chater, 1994), a form of rational analysis (Anderson, 1990). However, the details of such model are outside the scope of this paper (see Monaghan, Chater & Onnis, in preparation).

Overall, framing our results within a reduction of uncertainty principle should prompt new research aimed at discovering in which carefully controlled statistical environments multiple sources are attended to and either discarded or integrated.

Finally, our findings might inform research in language development. Gómez (2002) found that infants attend to the variability effect. We are currently investigating whether the U-shape curve found in our experiments applies to infant learning as well. The fact that performance in the zero-variability condition is very good is consistent with various findings that children develop productive linguistic knowledge only gradually building from fixed item-based constructions. According to the Verb Island hypothesis for example (for a review, see Tomasello, 2000) early knowledge of verbs and verb frames is extremely idiosyncratic for each specific verb. In addition, morphological markings are unevenly distributed across verbs. In this view *I-am-eat-ing* is first learnt as an unanalyzed chunk and it takes the child a critical mass of verbs to realize that the frame *am—ing* can be used productively with different verbs. Two- and three-year olds have been found to generalize minimally, their repertoire consisting of a high number of conservative utterances and a low number of productive ones. The speculation is that a critical number of exemplars is vital for triggering schematization. Perhaps then, young children exploit n-gram statistics as a default option, because their knowledge of language is limited to a few *type* items. This situation is similar to learning in small set sizes and it only works if each string is learnt as a separate item. When children's repertoire is variable enough (arguably at ages three to four), then switching to change versus non-change as a source of information becomes more relevant and helps the learner reduce uncertainty by detecting variant versus invariant structure. Although our experiments do not test for generalisation, the fact that learners in the large set size discard the middle item could be interpreted as a form of generalisation for material in the middle item position. At this stage the link between AGL results and language learning can only be speculative, but invites to intringuing research for the immediate future.

Acknowledgments

Luca Onnis and Nick Chater were supported by European Union Project HPRN-CT-1999-00065. Morten Christiansen was supported by Human Frontiers Science Program. Rebecca Gómez was supported by Grant NIH RO1 HD42170-01.

References

Anderson, J. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum Associates.

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

Dulany, D.E., Carlson, R.A., & Dewey, G.I. (1984). A case of syntactical learning and judgement: How conscious and how abstract? *Journal of Experimental Psychology: General*, 113, 541-555.

Fiser, J., & Aslin, R.N. (2002). Statistical learning of higher-order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 130, 658-680.

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.

MacKay, D.J.C., (1992). Information-based objective functions for active data selection. *Neural Computation*, **4**, 589-603.

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

McClelland, J.L. (1998). Connectionist models and Bayesian inference. In M.Oaksford, & N. Chater (Eds.) *Rational models of cognition*. Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4)**,** 608-631.

Onnis, L., Destrebecq, A., Christiansen, M., Chater, N., & Cleeremans, A. (submitted). *The U-shape nature of the Variability Effect: a connectionist model*.

Monaghan, P., Chater, N., & Onnis, L. (in preparation). Optimal data selection in sequential AGL learning.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, 119, 264-275.

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

Saffran, J.R., Johnson, E.K., Aslin, R.N., and Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition,* 70**,** 27-52.

Schvaneveldt, R.W., & Gómez, R.L. (1998). Attention and probabilistic sequence learning. *Psychological Research*, *61*, 175-190.

Servan-Schreiber, E., & Anderson, J.R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 592-608.

Tomasello, M. (2000). The item based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4, *156-163.*

Vokey, J.R. & Brooks, L.R. (1992). Salience of item knowledge in learning artificial grammar. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 328-344.