Grammar Induction Profits from Representative Stimulus Sampling

Fenna H. Poletiek (poletiek@fsw.leidenuniv.nl)
Department of Psychology, Leiden University, pobox 9555
Leiden, The Netherlands

Nick Chater (n.chater@ucl.ac.uk)
Department of Psychology, University College London, Gower Street,
London, WC1E6BT, UK

Abstract

Sensitivity to distributional characteristics of sequential linguistic and nonlinguistic stimuli, have been shown to play a role in learning the underlying structure of these stimuli. A growing body of experimental and computational research with (artificial) grammars suggests that learners are sensitive to various distributional characteristics of their environment (Kuhl, 2004; Onnis, Monaghan, Richmond & Chater, 2005; Rohde & Plaut, 1999). We propose that, at a higher level, statistical characteristics of the full sample of stimuli on which learning is based, also affects learning. We provide a statistical model that accounts for such an effect, and experimental data with the Artificial Grammar Learning (AGL) methodology, showing that learners also are sensitive to distributional characteristics of a full sample of exemplars.

Keywords: Artificial grammar learning; statistical learning; frequency distribution

Introduction

People seem naturally sensitive to structural characteristics of their environment, and they are able to use this knowledge adaptively. Such learning occasionally occurs implicitly and without instruction. For example, learning motor patterns like tying shoe laces and riding a bicycle, and several aspects of social behavior involve implicit associative learning. However, learning the rules of language by children is probably the most striking example of acquiring structure knowledge without apparent explicit awareness. Though it is currently debated to what extent an innate predisposition is responsible for this achievement or a general inductive learning capability, a growing number of studies suggests that humans have a powerful and adaptive sensitivity to the statistical properties of environmental stimuli (Kuhl, 2004; Gomez & Gerken, 2000; Redington, Chater & Finch, 1998).

In particular, studies on implicit sequence learning have revealed that statistical patterns can be picked up and used in subsequent usage of the system. For example, distributional characteristics in linguistic materials have been suggested to support syntactical category learning (Onnis, Monaghan, Richmond, & Chater, 2005; Mintz, 2002). Infant studies suggest that babies segment a stream of sounds (artificial words and syllables), on the basis of statistical associative properties of these sequences (Saffran, Aslin & Newport, 1996). Recently, another kind of regularity in sequential information has been proposed to affect segmentation: i.e. differences in variability between adjacent elements. For example, the word 'walk*ing'* consists of the highly variable part: (walk) and the invariant part 'ing'. This difference in variability has been proposed to serve as a cue for finding the boarders of linguistic units such as words (Gomez, 2002; Monaghan, Onnis, Christiansen & Chater, submitted). Finally, semantic regularities and associations are proposed to play a role in learning grammatical regularities (Rohde & Plaut, 1999): some words are much more associated than others, for semantic reasons. E.g., *he walks* is more frequent than *'he city'*, suggesting permissible and impermissible word (category) order.

Besides the statistical characteristics regarding local transitions in a structured sequence, statistical characteristics of the full stimulus sample of exemplars with which a learner is trained, may be informative for grammar induction as well (Poletiek, 2006). Consider a grammar G producing exemplars varying in length. A random output of such a grammar would be a sample containing short and long exemplars exemplifying all kinds of sequential rules specified by the grammar, but not all rules in equal number. Indeed, such output would contain more exemplars exemplifying typical and highly frequent rules in the grammar, (for example, highly associated adjacent or non adjacent elements) than exemplars with exceptional rules (Chater & Vitányi, submitted). Also, in a general case, short exemplars would occur more often than long ones. The resulting frequency distribution of this random output sample of G, may provide information to a learner for inducing G. Thus, in addition to the distributional features (e.g., sounds, syllables and words in natural language) *within* exemplars, the learner may benefit from distributional characteristics *between* exemplars of a full input sample (Poletiek, 2006).

In natural language, these high level distributional characteristics of the linguistic input are obvious. Some grammatical constructions are much more frequent than others. For example, sentences with three (levels of) self embedded relative clauses are rare as compared to sentences

with no self embedding or sentences with the sequence: noun-verb-object, which occur often in every day spoken language. If short sentences, and sentences with typical rules and transitions, indeed occur more often in the output of a given language then we may expect this kind of exemplars to be overrepresented in input samples directed at children.

Studies on child directed speech have indeed shown that the sample of utterances a child is faced with systematically differs from adult speech. Child directed speech (CDS) indeed includes more short sentences and typical sentences, and fewer complex sentences and subordinate clauses. Also, it contains more exactly repeated sentences (Snow, 1972; Philips, 1973; Pine, 1994). Whether these differences between child and adult directed speech in natural language is functional for natural language learning and in what sense, is still under debate (Gallaway & Richards, 1994). We propose that the frequency variations between different kinds of exemplars in a random output sample generated by a grammar, may not be vain but informative in the learning process.

In the present study, we propose an account for and a test of the learner's sensitivity to higher order statistical characteristic of the stimulus input. We use the Artificial Grammar Learning paradigm to test the hypothesis that the statistical properties of an input set are facilitative to the learning. Specifically, we propose that when the unique exemplars of a sample are distributed in accordance with their probability to be generated by the grammar, learning is facilitated.

Artificial Grammar Learning

In Reber's (1967) now classic experimental paradigm of Artificial Grammar Learning (AGL), implicit induction of sequential structure knowledge has been evidenced. In the first phase of the standard task, participants are presented with a number of exemplars from an artificial Finite State grammar (a Markov grammar), without being informed about the grammatical system. Next, participants are informed that the exemplars satisfied a rule system. In the second phase of the task, they categorize new sequences as correct or incorrect according to the grammar. Typically, participants perform better than chance, though they are unable to tell what exactly they learned. This is often interpreted as evidence for implicit learning of the grammar; although authors disagree about what knowledge is precisely induced. Better than chance performance in AGL, may either be explained by learning the *rules* of the grammar, or by (e.g., Redington & Chater, 1996), mere encoding of local statistical *patterns* in the sequences.

Statistical learning approaches of grammar induction in AGL propose that participants learn multiple kinds of regularities within the exemplars, and that they endorse new strings that exhibit those regularities at test. Participants might learn the frequencies of bigrams in a training set and endorse new strings that exhibit those bigrams (Perruchet & Gallego, 1997; Perruchet & Pacteau, 1990), but local transitional probabilities (Poletiek & Wolters, submitted) and variability transitions have also been shown to be involved in AGL (Onnis et al., 2004).

We hypothesize that the statistical properties of the whole input set of stimuli contribute to performance in AGL. In particular, the differential occurrences are a useful cue to the learner of the grammar. In the following, this effect is accounted for. We first present the formal model specifying the probabilities of each unique exemplar to be generated by the grammar. This model allows deriving the frequency distribution of a random output of a typical Finite State grammar used in AGL. Second, experimental data are presented that compare an AGL learning condition with equally distributed exemplars, to a learning condition with exemplars unequally distributed according to their probabilities to be generated by the grammar.

The Probability Distribution of Exemplars of G

The probability of an exemplar to be generated by a grammar can be calculated as the product of the path probabilities it 'follows' through the grammar. In Figure 1, the scheme of grammar G (Meulemans & van der Linden, 1997) is displayed, with the path probabilities added (see also Poletiek, 2006).

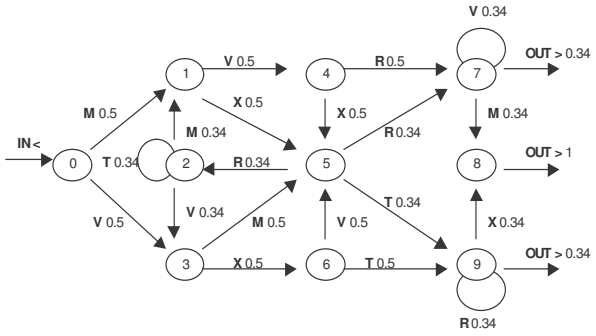

Figure 1: Finite State Grammar G used in previous (Meulemans & van der Linden, 1997) and the present AGL study.

The values represent the probabilities that a path is 'taken' when starting from the state it issues from. In Figure 1, these probabilities are unbiased. That is, all paths issuing from the same state are equally probable. The probability of an exemplar to be produced by G, is the product of these path probabilities. For example, the exemplar MVRM's probability to be generated by G is p(MVRM|G) = .5x.5x.5x.34 = .0416

Notice that the value p(exemplar|G) varies according to two dimensions: the average probabilities of the paths it takes and the length of the exemplars. Indeed, as a string is shorter, and therefore the product of the path probabilities contains less probabilities to be multiplied, the product increases. The probabilities of the unique exemplars of G

represent their frequency distribution in a random output of G: high p-value exemplars will be generated proportionally more often than low p-value (rare) exemplars (Chater & Vitányi, submitted).

Notice that the p's nicely model the characteristics of child directed speech. Child directed speech is characterized by short utterances and simple typical grammatical rules (high probability transitions), corresponding in our model to high p exemplars.

In the following study, we test whether presenting exemplars in accordance with their frequency distribution in a random output of the grammar, helps to induce this grammar.

Experiment

One group of participants in an artificial grammar learning task was exposed to a learning set of exemplars of G (Figure 1) distributed in accordance with their probabilities to be generated by the grammar. Another group was exposed to the same unique exemplars, all presented an equal number of times: i.e. in a 'flat' frequency distribution. We hypothesize that the distributional information in the stimulus set of the first group is used by the learner as a cue for making sense of the underlying rules. Presumably, it allows the learner to separate 'basic' rules or most typical rules in G from rare, exceptional complex paths. This variability, we suggest, supports the inductive learning process by suggesting priorities in rules learning. Hence, higher performance is expected in the unequal frequencies than in the equal frequencies group.

Method

Participants 24 students of Leiden University participated in the experiment on a voluntary basis.

Materials A sample of 34 unique exemplars from G was generated. These 34 exemplars were presented about 100 times in total during training. In one condition, each unique exemplar was presented 3 times (the equal frequencies condition), in the other condition, each exemplar was presented one or more times (varying from one to seven times) depending on its p-value. The appendix displays the training items, their p-value in G (p(exemplar|G)), and their frequency of occurrence in the training set under both conditions. In the equal frequencies condition, the total number of stimuli at training sums up to 102, being a plural of 3 (see Appendix).

The test set was made of 32 strings, half of which were grammatical and half (16) ungrammatical.

Procedure Participants were run in groups of four to six. The stimuli were displayed on a video screen. Participants were randomly assigned to the learning conditions (12 to the Equal frequencies condition; 12 to the Unequal frequencies condition). The instructions were in line with the implicit learning instructions in a standard AGL procedure (Reber, 1976). Participants were told they were participating in a memory experiment and would see letter strings. Their task was to remember them as well as possible. The order of presentations of the hundred strings was randomized. A group of four strings was shown for ten seconds. Afterwards, the participants were encouraged to write down the strings, from memory, to help memorizing. The strings were shown again for three seconds and after a blank of two seconds a new set of strings appeared.

At the beginning of the test phase, the participants were told that there were rules determining the sequence of letters in the strings they just tried to memorize. Participants were further told that they would see new strings, some of which followed these rules, while others did not. These strings would appear one by one, during a period of four seconds, followed by a blank of three seconds. Their task was to classify each string as correct or incorrect according to the rules of the learning phase. A response form was provided, which they could use to indicate their answers. All groups received the same test items and the same test instructions.

Results

The mean number of correct categorizations was 15.3 (48%) (N=12; sd=3.8) for the participants trained with an equal distribution. Participants trained with an unequally distributed learning set of exemplars was 20.1(63%) (N=12; sd=1.2). A t-test for independent samples (equal variances not assumed) was significant in the expected direction (t= -4.1;df=13.3; p=.001).

Discussion

The participants trained with exemplars, unequally distributed in frequencies according to their probabilities to be produced in a random generation of the grammar, learned better than participants trained with the same unique exemplars of G distributed equally. The effect was large. This result is in line with our hypothesis that learners exploit the distributional characteristics of unique exemplars in the learning input to induce the underlying structure. The frequency distribution of occurrence of exemplars they are faced with, serves as a facilitative cue to make sense of the system producing the exemplars, as we predicted. What cognitive mechanism may underlie this effect of high order statistical properties? In what sense is the frequency distribution facilitative? Two links with previous findings are explored.

Possibly, the G-representative distribution helps the participant to learn fragments, i.e. bigrams. In other words, higher order statistical characteristics may be helpful to learn lower order characteristics. Indeed, if participants proceed by learning distributional information about bigrams during training, then a distribution of exemplars that is valid for the full language that is naturally generated by the grammar, will also give more valid distributional information about the occurrence of the fragments in the full output of the grammar. This was shown in a simulation study (Poletiek, 2006) in which the fragment knowledge of two learners were compared; one trained with a flat and

another with a grammar-representative distribution of exemplars. Hence, if the learning strategy is fragment based, then the unequal frequency distribution of the learning set provides the learner with more valid information about the occurrence of fragments in the grammar, which in turn can be expected to help for recognizing new stimuli satisfying the structure (Poletiek & Wolters, submitted).

Another possible perspective on the present frequency distribution effect, relates to the recent proposal stressing the importance of variability within the structured stimuli on learning regularities (Onnis et al., 2004). In this proposal, differences in variability of both adjacent and non-adjacent elements in exemplars of the language have been shown to help the learner to associate elements, and segment streams of elements, into new unities. It seems that the differences in frequencies and variance of consecutive elements in a stimulus input play an important role in structure induction.

Our results suggest that variability may play a role at a higher level as well: Differences in frequencies of occurrence between stimuli make it possible to differentially strengthen sequential rules, rather than learn all rules equally well without any priorities. Variability in frequencies of occurrence of stimuli may tune the learner to hierarchical learning. Indeed, if the frequently occurring exemplars of the structure are simple and more typical for the system than low frequent ones, then the basic grammatical constructions may be very thoroughly acquired due to their frequent presentation. Oppositely, more exceptional rules exemplified in less frequently seen stimuli, may be mastered more superficially and acquired in a later stage of exposure. Variability, thus, may play a rather fundamental role at several levels of processing the input of structured sequential materials.

Though we could verify the predicted main effect of frequency distribution in the present AGL study, a surprising observation was the poor performance by the participants in the condition with a flat frequency distribution. We explore a number of explanations. First, the low performance in the condition in which participants saw each exemplars repeated equally often may reflect a possible negative effect of the lack of variability. In other words, not only may variability in input have a positive effect on learning, lack of variability may hamper the inductive learning process. Recent data from our lab suggest that lack of variability may be more detrimental for a learner than one may intuitively expect: participants presented with a sample of exemplars of a grammar (potentially generating strings from differing lengths) all having the same length, unexpectedly showed no learning.

A second tentative explanation is that the instruction to memorize the exemplars, though facilitating the recognition of common rules and patterns of the grammar between exemplars in the unequal frequencies condition, has focused attention on the unique characteristics of individual exemplars equal frequencies condition, deviating attention from what the exemplars have in common. Thus, perhaps these participants actually memorized very well *individual* items, but this might have interfered with picking up implicitly the structure common to all of them. Since the participants were tested on completely new items and a grammaticality judgments task, the memory of every individual item was of no advantage.

Pushing this line of reasoning a bit further, a relation may be sketched between the memorize task at training, the grammaticality judgments task at test, and the stimulus distribution. Unequal frequencies (rather than equal frequencies) conditions may specifically facilitate rule induction, if the unequal distribution models the grammar's output, but not necessarily affect memorization (recall or recognition) *performance*. Suppose that the learner learns to discriminate between grammatical and ungrammatical items by building an approximate probabilistic *model* of the underlying distribution. We do not need to make any particular assumptions about the nature of this approximation -- it might be that the model is based on induced rules, or bigrams, or transitional probabilities or information of any other type. This probabilistic model will be likely to be a good approximation to the true distribution, if the frequencies of the stimuli correspond to their relative probabilities according to the underlying generating process (the Markov grammar). By contrast, to the extent that the items are atypical, the underlying probability distribution should be difficult to learn (for relevant formal analysis of typicality and learnability, see Vitányi & Li, 2000). In particular, then, if items are repeated *evenly* rather than according to the generating Markov distribution, we might expect learning of the probability distribution to be impaired.

A different pattern of predictions arises, however, if we assume that learning is purely determined by memory for the individual items seen during training. Indeed, mere *repetition* of stimulus items during memorization, has been shown to facilitate *recollection* in early memory research (Ebbinghaus, 1913), or as suggested by studies about presentation time effects (Roberts, 1972; Waugh, 1967). Hence, assuming rote memory during training, we may expect either equal performance in both distribution conditions (if we assume the process to be linear) or better performance in the equal distribution condition (if we assume the process to be convex) on a pure recollection or recognition task. Of course, the implicit grammar learning paradigm is interested in structure learning rather than memory performance. Therefore, learners are tested on their knowledge of the grammar by means of a categorization task, yet after having been instructed to *memorize* the items at training (Reber, 1976).

In sum, the poor performance on the grammaticality judgments task in the equal frequencies condition may be related to interference between the task carried out at training and the task carried out at test. A stimulus distribution that is helpful for memorization may not be for expressing grammar knowledge. Our research currently investigates this idea further.

In conclusion, distributional sample characteristics of structured input, seems to affect structure learning in a positive way. This is in line with learners' sensitivity to lower order statistical regularities, which have been amply documented in the literature. Also, it catches the observation

in natural language acquisition, that young language learners are exposed more often to very typical and short linguistic input in the period in which they learn the language. In addition, the present study points at the relevancy of simple experimentation with artificial stimulus materials starting from parsimonious assumptions, to understand our sensitivity to sample characteristics of environmental stimuli, and how we exploit it to extract useful structure knowledge.

Acknowledgments

We would like to thank Gordon Brown for his input on the relation between repetition and performance in rote memory. Correspondence can be sent to Fenna Poletiek poletiek@fsw.leidenuniv.nl

References

Chater, N.& Vitanyí, P.(submitted). A simplicity principle for language acquisition. Re-evaluating what can be learned from positive evidence.

Ebbinghaus, H. (1885). *Über das Gedächtnis*. Leipzig: Dunker. (Translation by H. Ruyer and C.E Bussenius (1913). Memory. New York: Teachers College)

Gallaway, C., & Richards, B.J. (Eds.) (1994) *Input and interaction in language acquisition*. Cambridge: Cambridge University Press.

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

Gómez, R.L. & Gerken, L.A. (2000). Artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*, 178-186.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews: Neuroscience, 5,* 831-843.

Meulemans, T., & van der Linden, M. (1997). Associative chunk strength in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23* , 1007-1028.

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

Monaghan, P., Onnis, L., Christiansen, M., & Chater. N. (submitted). The importance of being variable: Learning nonadjacent dependencies in speech processing.

Onnis, L., Monaghan, P., Christiansen, M.H., & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. *Proceedings of the 26th Conference of the Cognitive Science Society*. Mahwah, NJ:Lawrence-Erlbaum.

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, *53*,225-237

Perruchet,P. & Gallego, J. (1997). A subjective unit formation account of implicit learning. In D.C.Berry (Ed), *How implicit is implicit learning? Debates in psychology* (pp. 124-161). Oxford , England:Oxford University Press.

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

Philips, J. R. (1973). Syntax and vocabulary of mothers' speech to young children: age and sex comparisons. *Child Development, 44,* 182-185.

Pine, J. M. (1994). The language of primary caregivers. In Gallaway, C.,& Richards, B.J. (Eds.) *Input and interaction in language acquisition*. Cambridge: CUP

Poletiek, F.H. (2006). Representative sampling in an artificial grammar learning task. In P.Juslin, & K.Fiedler (Eds). *Information sampling and adaptive cognition* (pp. 440-455). Cambridge: Cambridge University Press.

Poletiek, F.H., and Wolters, G. (submitted). A model for grammaticality and chunk associativeness in Artificial Grammar Learning.

Reber, AS (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 6, 855-863.

Reber, A.S. (1976). Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 88-94.

Redington, M. & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General*, *125*, 123-138.

Redington, M., Chater, N.,& Finch, (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425-269.

Roberts, W.A. (1972) Free recall of word lists varying in length and rate of presentation: a test of total-time hypotheses. *Journal of Experimental Psychology, 92*, 365-372.

Rohde, D.L.T., & Plaut, D.C. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition 72*, 67-109.

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

Snow, C.E. (1972). Mothers' speech to children learning language. *Child Development, 43*, 549-65.

Vitányi, P. M. B., & Li, M. (2000). Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity, IEEE Transactions in Information Theory, IT-46, 446-464.

Waugh, N.C. (1967) Presentation time and free recall. *Journal of Experimental Psychology. 73*, 39-44.

Appendix

Training stimuli with p-values and frequencies (f(ex.)) of presentation in the unequally distributed learning set condition, and the equally distributed learning set condition.

Exemplar	p(ex.|G)	f (ex.) (uneq.)	f (ex.) (eq.)
MVR	0.04166	7	3
MVRM	0.04166	7	3
MVRV	0.01388	3	3
MVXRVM	0.00462	2	3
MVXRVVV	0.00051	1	3
MVXTX	0.01388	3	3
MXR	0.02777	5	3
MXRM	0.02777	5	3
MXRMXRVMM	0.00051	1	3
MXRTMXRVXTX	0.00004	1	3
MXRV	0.00925	3	3
MXT	0.02777	5	3
MXTR	0.00925	3	3
MXTRX	0.00925	2	2
VMR	0.02777	5	3
VMRMVRV	0.00077	1	3
VMRV	0.00925	2	3
VMRVM	0.00925	3	3
VMT	0.02777	4	3
VMTRR	0.00308	2	3
VMTRRR	0.00102	1	3
VMTX	0.02777	5	3
VXT	0.04166	7	3
VXTR	0.01388	3	3
VXTRR	0.00462	2	3
VXTRRX	0.00462	2	3
VXTRX	0.01388	3	3
VXTX	0.04166	2	3
VXVR	0.01388	3	3
VXVRM	0.01388	3	3
VXVRTMXRV	0.00008	1	3
VXVRTTVXVRVXVTR	0.000007	1	3
VXVRVMRMXRMVR	0.000003	1	3
VXVTRR	0.001543	1	3