Individual behavior in learning of an artificial grammar

Vitor C. Zimmerer · Patricia E. Cowell ·
Rosemary A. Varley

Published online: 17 December 2010
© Psychonomic Society, Inc. 2010

Abstract Artificial grammar learning (AGL) is a widely used experimental paradigm that investigates how syntactic structures are processed. After a familiarization phase, participants have to distinguish strings consistent with a set of grammatical rules from strings that violate these rules. Many experiments report performance solely at a group level and as the total number of correct judgments. This report describes a systematic approach for investigating individual performance and a range of different behaviors. Participants were exposed to strings of the nonfinite grammar A^nB^n . To distinguish grammatical from ungrammatical strings, participants had to pay attention to local dependencies while comparing the number of stimuli from each class. Individual participants showed substantially different behavioral patterns despite exposure to the same stimuli. The results were replicated across auditory and visual sensory modalities. It is suggested that an analysis that looks at individual differences grants new insights into the processes involved in AGL. It also provides a solid basis from which to investigate sequence-processing abilities in special populations, such as patients with neurological lesions.

Keywords Syntax · Artificial grammar learning · Individual behavior · Modality

V. C. Zimmerer
Abteilung für Allgemeine Sprachwissenschaft,
Heinrich-Heine-Universität Düsseldorf,
Universitätsstr. 1,
40225 Düsseldorf, Germany

P. E. Cowell · R. A. Varley (✉)
Department of Human Communication Sciences,
University of Sheffield,
31 Claremont Crescent,
Sheffield S10 2TA, UK
e-mail: r.a.varley@shef.ac.uk

Artificial grammar learning (AGL) is an experimental paradigm that investigates pattern detection in syntactically structured stimuli. The typical AGL experiment (Reber, 1967) consists of two phases: In a training phase, participants encounter stimulus strings that are generated by a set of grammatical rules (the target grammar). In a subsequent test phase, participants encounter new stimulus combinations. Some are consistent with the target grammar, while the others violate it. Participants are asked to accept or reject the test strings on the basis of their experience during the training phase. A high rate of correct decisions is claimed to reflect spontaneous syntactic learning. The paradigm allows control of several cognitive processes that might not be of primary interest. The stimuli can be nonwords and nonsensical, which reduces demands on lexical and semantic processing. It is also possible to reduce or eliminate demands on phonological processing by using short, phonologically simple items or by switching to a different sensory modality altogether.

It is assumed that AGL helps one to understand the cognitive processes that underlie natural syntactic processing. AGL data are discussed in a range of different linguistic frameworks (e.g., Goldberg, 2006; Hauser, Chomsky, & Fitch 2002; Tomasello, 2003). AGL experiments have been performed to investigate syntactic pattern detection in infants (Marcus, Vijayan, Bandi Rao, & Vishton 1999; Saffran, 1996; Saffran, Johnson, Aslin, & Newport 1999), monkeys (Fitch & Hauser, 2004; Hauser, Newport, & Aslin, 2001; Saffran et al., 2008), starlings (Gentner, Fenn, Margoliash, & Nusbaum 2006), and pigeons (Herbranson & Shimp, 2003). AGL has also been used to explore syntactic cognition in patient populations, such as people with aphasia (Caplan, Baker, & Dehaut 1985; Hoen et al., 2003), Parkinson's disease (Smith, Siebert, McDowall, & Newport 2001), and dyslexia

(Rüsseler, Gerth, & Münte 2006), in order to determine the integrity of sequential pattern detection abilities. It has further been used to map the neural systems responsible for certain syntactic processes, such as monitoring transition probabilities (Petersson, Forkstam, & Ingvar 2004; Uddén et al., 2008) or hierarchical embedding (Bahlmann, Schubotz, & Friederici 2008).

In this study, we explored the question of whether individual participants who have been exposed to the same stimuli acquire different knowledge and, thus, apply different decision criteria in the testing phase. This question has been identified as “underresearched” (Visser, Raijmakers, & Pothos 2009). For instance, all of the studies mentioned above looked primarily at group data, and many focused on the question of whether the target grammar was “mastered” or not. A high mean percentage of correct judgments was seen as evidence for the participants’ acquisition of the target grammar, or at least a grammar that was very similar to it. However, it is possible to achieve high performances without mastering the target grammar. A number of studies have shown that fragmentary knowledge, such as knowing how a grammatical string can start or end or which chunks (bigrams, trigrams) can appear, is often sufficient to explain above-chance performance in the test phase (Dulany, Carlson, & Dewey 1984; Perruchet & Pacteau, 1990; Redington & Chater, 1996). There is also evidence that participants base their decisions on a Gestalt approach, determining how similar the test strings are to the training strings they encountered (Pothos & Bailey, 2000). If a range of successful decision criteria is possible, it is likely that different individuals come to different representations of syntactic knowledge, even if the stimuli are the same for all.

The analysis of group data, as seen in the majority of AGL reports, provides a picture of the *average learner*. Although these studies report measures of statistical dispersion, they cannot fully account for individual differences. In the worst case, group data describe a behavioral pattern that does not occur within a single individual in the group. Investigating individual performance is critical for a more in-depth understanding of the processes involved in AGL. This might be important in the context of functional imaging studies, where different neural mechanisms correspond to different processing patterns. The analysis of individual behavioral patterns is particularly salient for the analysis of data from intrinsically heterogeneous groups, such as people with brain damage. Since no two brain lesions are anatomically the same and there may be individual differences in brain organization and learning, case, rather than group, studies are frequently used. To find out whether one individual pattern is atypical, a description of the typical patterns is required to establish normative

baselines. Previous patient work, however, has only compared patient with control groups, without looking at individual behavior (Caplan et al., 1985; Hoen et al., 2003).

Visser et al. (2009) proposed a method for exploring individual AGL performance based on regression models. Regression models had been used previously to examine individual behavior in the balance scale task (Jansen, Raijmakers, & Visser 2007; van der Maas & Straatemeier, 2008). Visser et al. reanalyzed data from previous AGL experiments to determine whether individuals were more likely to base their decisions on grammaticality (whether test strings were generated by the target grammar) or fragment overlap (how similar they were to training phase exemplars). They discovered that individuals indeed applied different decision criteria. For example, in Pothos, Chater, and Ziori’s (2006) experiment using a finite-state grammar and letters as stimuli, 10 of the 20 participants based their decisions predominantly on grammaticality, 8 on fragment overlap, and 2 on neither criterion. Visser et al. demonstrated how the analysis of individual behavior can provide new insight into learning processes. Instead of describing the average learner, they identified grammaticality judgments and similarity judgments as two distinct cognitive processes in AGL.

The precision of regression models increases with the size of the data set, so individual behavior can be better described when the samples are large. This article presents a different approach that allows the description of single cases independently of sample size. This was made possible by increasing the number of test trials for each participant. While test phases in AGL experiments have usually consisted of a few dozen trials (the experiments reported in Visser et al., 2009, had 32), the two experiments designed for this study had 120 test strings. The aim of the present study was to assign a learning profile to each of the participants and, ultimately, to find out how performance patterns within the group could be described on the basis of individual performances. The length of the test phase also allowed for exploration of changes in behavior within the experiment, since it was possible that participants changed their decision criteria across the course of the test phase.

Grammatical strings in both experiments were generated by the nonfinite grammar A^nB^n , which has been widely discussed in the recent literature (de Vries, Monaghan, Knecht, & Zwitserlood 2008; Fitch & Hauser, 2004; Friederici, 2004; Hochmann, Azadpour, & Mehler 2008; Perruchet & Rey, 2005). In A^nB^n strings, a number of items of class A are followed by the same number of items of class B (e.g., AB , $AABB$, $AAABBB$, etc.). In one of the first AGL experiments using the grammar, Fitch and Hauser exposed participants to A^nB^n strings, using spoken syllables as stimuli. Class A syllables were produced by a female

speaker, and class B syllables by a male speaker. Maximal string length in the training phase was six syllables. After exposure to 30 grammatical strings, participants were able to successfully distinguish grammatical from ungrammatical strings in the test phase (85% of grammaticality judgments were correct). Participants were also able to generalize what they learned to strings of eight syllables.

It was initially assumed that participants acquired the grammar by using recursive phrase-structure rules (Fitch & Hauser, 2004; Friederici, 2004; Hauser et al., 2002). A different account is now more widely accepted. Participants who acquired the grammar may have checked to see whether a block of *As* preceded a block of *Bs* and may have counted the number of items in each block to see whether they were the same (Bahlmann et al., 2008; Christiansen & Chater, 1999; de Vries et al., 2008; Hauser, Barner, & O'Donnell 2007; Perruchet & Rey, 2005; Pinker & Jackendoff, 2005). The task would thus require the detection of local dependencies (which elements can appear where), as well as the monitoring of numerosity.

The findings reported by Fitch and Hauser (2004) could not explore this possibility, since the ungrammatical strings were of the type $(AB)^n$ (e.g., *ABABAB*) or B^nA^n (e.g., *BBBAAA*). Participants could have rejected any string containing small chunks of stimuli that did not appear in the training phase, such as the bigram *BA*. This alternative behavior is sufficient to achieve a perfect score. De Vries et al. (2008) found that when counting violations (e.g., *AABBB*) were introduced into the test phase, their participant group rejected them significantly less often than bigram violations (e.g., *AABABB*). Hochmann et al. (2008) observed that participants consistently accepted counting violations, concluding that only transition patterns were learned.

In the two experiments reported here, participants were presented with strings generated by the target grammar A^nB^n . In the subsequent test phase, stimuli consisted of

grammatical strings, as well as five different types of violations:

- Type 1 (counting; e.g., *AABBB*) had an illegal total number of items in the string and did not have the same number of *As* and *Bs*.
- Type 2 (counting; e.g., *AAAABB*) did not have the same number of *As* and *Bs*.
- Type 3 (B^nA^n ; e.g., *BBBAAA*) had an incorrect initial and final item and contained the illegal bigram *BA*.
- Type 4 (block consistency; e.g., *BABAAB*) had an illegal initial item, contained the illegal bigram *BA*, and violated block consistency.
- Type 5 (block consistency; e.g., *ABBABA*) had an illegal final item, contained the illegal bigram *BA*, and violated block consistency.

Consistent rejection of all violation types and acceptance of grammatical strings would suggest the learning of A^nB^n . However, if individual participants fail to reject all violations and display different sensitivity to the violation types, this might indicate acquisition of different knowledge. Eight possible behavioral patterns that might be observed after exposure to the A^nB^n strings were identified. Each possible pattern was assigned one *template* for analysis. Templates were defined by the types of violations that would be rejected (Table 1). Each produced one score. For example, a participant who rejected only strings with the bigram *BA* would achieve a low score under the A^nB^n template, but a higher A^+B^+ score. The template under which the participant achieved the highest score was deemed the *best template* and, as such, represented the grammar most likely acquired by the individual.

Template labels were descriptive terms, with no claim regarding their underlying psychological reality. The observed behavior could be the result of rule acquisition, but also of processes based on associative learning or

Table 1 Behavioral templates for the analysis of A^nB^n experiments

Template	Description	Types of violations to reject	Strings to accept
A^nB^n	A number of <i>As</i> is followed by an equal number of <i>Bs</i>	1, 2, 3, 4, 5	A^nB^n
EVEN	Total number of items in a string is even	1	2, 3, 4, 5, A^nB^n
EQUAL	Equal number of <i>As</i> and <i>Bs</i>	1, 2	3, 4, 5, A^nB^n
$A(A+B)^*$	String starts with <i>A</i>	3, 4	1, 2, 5, A^nB^n
$(B+A)^*B$	String ends with <i>B</i>	3, 5	1, 2, 4, A^nB^n
A^+B^+	All <i>As</i> precede the <i>Bs</i>	3, 4, 5	1, 2, A^nB^n
$(A^+B^+)(B^+A^+)$	All items of one class precede all items of the other class	4, 5	1, 2, 3, A^nB^n
$(A^nB^n)(B^nA^n)$	All items of one class precede all items of the other class, equal number of <i>As</i> and <i>Bs</i>	1, 2, 4, 5	3, A^nB^n

Notation according to Cohen (1991): A^n , *n* instances of *A*; A^* , zero or more instances of *A*; A^+ , one or more instances of *A*; $A + B$, either *A* or *B*

judgments based on the similarity of a novel string to specific exemplars from the training phase (Pothos, 2007).

The two experiments differed by stimulus modality. The stimuli in [Experiment 1](#) were spoken nonword syllables. In [Experiment 2](#), they were nonsense visual stimuli. The development of parallel experiments across modalities facilitates work with patient populations. For example, patients with a particular impairment of phonological processing, as in the case of aphasia, might be tested in a visual modality. If performances (including individual behavioral patterns) of healthy participants are similar across sensory modalities, experimenters can select the modality of the test on the basis of the capacities of the patient participants.

In [Experiment 1](#), the effect of task instructions was also investigated. Some AGL experiments have explicitly asked participants to pay attention to patterns (e.g., Perruchet & Rey, 2005), whereas others used nonspecific instructions (Dulany et al., 1984). It is possible that the nature of task instructions has an effect on participant behavior. This question is also of relevance when investigating people with a language disorder, who might not understand the specifics of the instructions. In order to explore this possible effect, [Experiment 1](#) included one condition using specific and one using nonspecific instructions.

All the participants were native speakers of English with no reported history of neurological disorder. All the experiments had full approval from the University of Sheffield Research Ethics Committee. They were written and run using DMDX¹ and were presented on a Dell Inspiron 5150 laptop PC connected to stereo speakers.

Experiment 1

[Experiment 1](#) was designed as a partial replication of Fitch and Hauser (2004) and involved learning A^nB^n in the auditory modality, using a female speaker for syllables of class A and a male speaker for syllables of class B. The effect of specific versus nonspecific instructions was also examined.

Participants

Twenty participants with a mean age of 25 years (range, 18–42; 10 female, 10 male) were recruited. They were students and staff at the University of Sheffield. Participants were recruited via e-mail networks. They had no knowledge of the project prior to the experiment, were not familiar with the experimental paradigm, and were told that they were participating in an experiment about the “perception of sounds.”

Stimuli and procedure

The training phase consisted of 72 strings of the grammar A^nB^n , presented in three blocks, with a self-timed break between blocks. The value of n was 2, 3, or 4, so that each string was four, six, or eight syllables long. For each possible n , there were 24 strings. The test phase contained 60 strings constructed by the grammar A^nB^n and 60 violations of A^nB^n . There were 12 strings for each of the five violation types described in [Table 1](#). All strings were four, six, or eight items long, except for violations of Type 1, which were three, five, or seven items long. The test phase was again divided into three blocks, with a self-timed break between each block. The order of strings throughout the experiment was randomized using Microsoft Excel. In the test phase, the order was semirandom in that no more than four consecutive trials were either all grammatical or all ungrammatical. String presentation order was the same for each participant.

For each A and B , one of eight syllables was presented. Syllables were of lower frequency and were nonwords of a consonant–vowel structure that was consistent with English phonotactics. The stimuli of each class differed by speaker identity, as well as phonetic structure. The syllables of class A were /ʃʊ, fʊ, gʊ, mʊ, dæ, væ, kɛ, bɛ/ and were spoken by a woman. Syllables of class B were /ʌ, bʌ, tʌ, ʃʌ, zɒ, vɒ, ʒə, gə/ and were spoken by a man. Syllable frequencies were determined using the CELEX lexical database.² The mean frequency for the syllables of class A was 217 (range, 5–785), and of class B 230 (range, 11–789), per million tokens. Syllables were edited for equal length and intonation using Audacity.³ Syllable duration was 400 ms. A syllable never occurred twice in one string, and during the training phase, each syllable appeared the same number of times in each legal position. Within a string, there was a 200-ms pause between each syllable. Between each string, there was a pause of 4.5 s.

The experiment had two conditions that differed in terms of instructions, which were labeled *specific* and *nonspecific*. Written instructions were presented on the screen. In the nonspecific condition, the instructions before the training phase were the following: “Just relax and listen.” Before the test phase, they were the following: “You will now hear new sounds. Decide whether what you heard fits to the examples you heard before. Decision time is important, so decide as quickly as possible.” For the specific condition, the instructions before the training phase were the following: “You will now listen to sequences of sounds for a few minutes. Pay attention to the pattern of the sequences. You will later be asked to answer questions about these

¹ Available at <http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm>.

² Available at <http://www.ru.nl/celex>.

³ Available at <http://audacity.sourceforge.net>.

sequences.” Before the test phase, the instructions were the following: “You will now hear new sequences. They contain the same sounds you heard before. Decide whether the patterns of the new sequences follow the same principles as the sequences you heard before. Try to answer quickly, but accuracy is also important.”

The participants were tested individually, with the experimenter present in the room. Participants were first asked about their age, native language, and other languages they spoke. They were randomly assigned to either the specific or the nonspecific condition and were told to follow the instructions on the screen. The training phase took about 7 min. In the test phase, which was about 12 min long, participants had to decide whether the stimuli fitted those in the training phase or not. They made a decision after each string. The participants had 4.5 s to respond, and no feedback was given. Participants made their decision using a two-button mouse and pressed the green button to accept strings and the red to reject strings. The position of the green button (left or right) was counterbalanced across participants.

After the test phase, participants were asked an open question about whether there was anything to which they had paid attention in making the decision. When the answers were unclear, participants were prompted to clarify their response.

Results

The mean percentage of correct grammaticality judgments based on the criteria of A^nB^n was 75.2% ($SEM = 3.54$). Individual scores ranged from 51% to 100%. An analysis of variance with the between-subjects factors of instruction and gender revealed no significant main effects or interactions. The groups were pooled for all subsequent analyses.

Responses were analyzed by string type. Participants correctly accepted 77.43% ($SEM = 3.67$) of the strings constructed by A^nB^n . Performance varied across different types of violations of A^nB^n (Fig. 1). The percentage of rejections was 71.74% ($SEM = 4.46$) for Type 1, 68.75% ($SEM = 6.56$) for Type 2, 60.42% ($SEM = 6.25$) for Type 3, 78.6% ($SEM = 5.02$) for Type 4, and 83.3% ($SEM = 4.72$) for Type 5. An ANOVA with violation type as the repeated measure showed a significant main effect, $F(4, 76) = 5.879$, $p < .001$. Violations of Type 3 were rejected less often than violations of Type 4, $p = .002$, and Type 5, $p < .001$, remaining significant after Bonferroni correction (adjusted threshold: $p < .005$). Violations of Type 5 were rejected more often than violations of Type 1, $p = .019$, and Type 2, $p = .014$. Violations of Type 3 were rejected less often than violations of Type 1, $p = .047$. These effects did not withstand Bonferroni corrections.

Strong correlations were observable between the counting violations (Types 1 and 2), $r(18) = .709$, $p < .001$, and

Fig. 1 Rejection patterns for the five string types in the auditory condition. Bars represent percentages of correct rejections ($M \pm SEM$). Viol. 1: Illegal total number of items (e.g., *AABBB*). Viol. 2: Number of *As* and number of *Bs* mismatch (e.g., *AAAABB*). Viol. 3: B^nA^n (e.g., *BBBAAA*). Viol. 4: Illegal initial item, violation of block consistency (e.g., *BABAAB*). Viol. 5: Illegal final item, violation of block consistency (e.g., *ABBABA*). Viol., violation type; cons., consistency

the violations of block consistency (Types 4 and 5), $r(18) = .834$, $p < .001$. Due to the correlations, nonsignificant differences and structural similarities between some violation types, counting violations (Types 1 and 2) were pooled for further analysis, as were block consistency violations (Types 4 and 5). The mean performance for counting violations was 70.25% ($SEM = 5.11$), and for block consistency violations, it was 80.95% ($SEM = 4.66$).

A repeated measures ANOVA (significance threshold adjusted by Bonferroni correction: $p < .017$) showed significant differences between the pooled string types, $F(2, 38) = 7.818$, $p < .001$. Performance was significantly higher for block consistency violations than for counting ($p = .015$) and Type 3 ($p = .001$) violations. The correlation between counting and block consistency violation scores was $.67$ ($p = .001$). There was a positive correlation between Type 3 and block consistency violation scores, $r(18) = .60$, $p < .005$. There was no significant correlation between Type 3 and counting violation scores.

Template analysis

Individuals' responses were analyzed for the best-fit template (Table 1). Template scores were not calculated as the percentage of correct decisions, because except for template A^nB^n , the number of grammatical and ungrammatical items was different. For example, under the template A^+B^+ , 36 of 120 strings were ungrammatical, while under the template $(A^+B^+)(B^+A^+)$, there were 24

ungrammatical strings. Therefore, any *no* bias would have favored templates with fewer grammatical strings, and any *yes* bias would have favored templates with more grammatical strings. To control for such biases, template scores were calculated as *difference scores* or *D scores* (Perruchet & Pacteau, 1990, p. 267). *D scores* are calculated as the percentage of ungrammatical strings rejected minus the percentage of grammatical strings rejected. The scale of *D scores* starts at 100 and ends at 100. Zero values indicate that ungrammatical and grammatical strings were rejected at the same percentage. Each template yielded one *D score*. The highest scoring template was the one that best described the decision criteria of the individual participant. Participants' data were also analyzed for chance performances. Above-chance status was achieved when the percentage of rejected ungrammatical strings was above chance and the percentage of rejected grammatical strings was below chance. Chance levels were determined using binomial tests (see the Appendix). Therefore, a high *D score* was not sufficient for a performance to be classified above chance.

A *D score* analysis suggested that across individual participants, alternate grammars were acquired (Fig. 2). Out of the 20 participants, 7 showed responses based on the learning of A^nB^n . Four participants showed response patterns based on the grammar $(A^nB^n)+(B^nA^n)$, accepting strings in which block B preceded block A. One participant based responses on A^+B^+ , accepting counting violations. Two participants showed responses based on $(A^+B^+)+(B^+A^+)$, rejecting only violations of block consistency. The remaining

six performances did not meet the criteria for being above chance.

There was no apparent difference between participants who received specific instructions and participants who received nonspecific instructions. Participants from both groups scored above chance on the A^nB^n , $(A^+B^+)+(B^+A^+)$, and $(A^nB^n)+(B^nA^n)$ templates. The participant who achieved his highest score on A^+B^+ had received specific instructions. Three participants from each group scored at chance on all templates.

To determine whether participant behavior changed within the test phase, *D scores* for the first 60 trials were compared with the scores for the last 60 test trials within each participant. There was no indication of a change in grammaticality criteria. On the contrary, in the second half of the test phase, participants became more consistent at applying their criteria. For those participants who performed above chance, the average *D score* for the best template was higher in the second half (73.62) than it was in the first (66.56). A paired-sample *t* test revealed a trend toward significance, $t(13) = 1.682$, $p = .116$.

Participant self-reports

In post hoc analyses, participants were split into four groups on the basis of their self-reports: Participants who reported having no criteria or judging by intuition ($n = 3$), participants who reported listening for familiar sounds or combinations of sounds ($n = 3$), participants who reported local rules (e.g., that a string started with a female speaker and ended with a male, or that the syllables came in two blocks; $n = 2$), and participants who reported a combination of local rules and comparing the number of female-spoken syllables with the number of male-spoken syllables ($n = 12$). No participant reported comparing the number of syllables from each class without applying decision criteria based on local features of the stimuli. Table 2 shows that participants who scored highest under the A^nB^n template reported a combination of local dependencies and counting. None of the participants who reported having no decision criteria or using intuition consistently applied syntactic rules.

Discussion

The overall percentage of correct decisions given the criteria of A^nB^n (75.2%) was lower than that reported by Fitch and Hauser (2004; 85%). However, Fitch and Hauser's experiment did not fully explore the learning of A^nB^n , due to the absence of counting violations. It can be assumed that without counting violations, participants in Experiment 1 would have made a greater number of correct decisions. The present results are closer to those of de Vries

Fig. 2 Individual participant performance depicted as a function of best fit in the grammatical template analysis. See Table 1 for details of the eight possible grammatical templates. Single dots can correspond to several cases with identical values

Table 2 Self-reported decision criteria of individual participants in relation to the best grammatical template (numbers in parentheses represent chance-level performances)

Criteria reported after the task	Best Grammatical Template				
	A^nB^n	EVEN	A^+B^+	$(A^+B^+)+(B^+A^+)$	$(A^nB^n)+(B^nA^n)$
No criteria/intuition		(2)		(1)	
Familiar units		(1)	1		1
Local				1(1)	
Counting + local	7			1(1)	3

et al. (2008; 73%), who employed violations similar to those used in the present experiment.

Although the participants were generally able to acquire aspects of A^nB^n , performance differed between violation types. In particular, participants rejected strings of B^nA^n less often than other ungrammatical strings. The analysis of individual performances using grammatical templates revealed that only 7 of the 20 participants acquired the target grammar with some likelihood. Six showed behavior not significantly above chance for any template. The others acquired a subset of the rules that constitute A^nB^n . Most of the participants reflected their behavior correctly in self-reports, indicating explicit awareness of their learning. There was no significant effect of gender or instructions.

Experiment 2

In Experiment 2, the learning of A^nB^n was tested in a visual modality to see whether learning patterns would be similar across sensory modalities. Instead of spoken syllables, Experiment 2 employed abstract shapes. Only one set of instructions, based on the nonspecific instructions in Experiment 1, was used.

Participants

Twenty participants with a mean age of 27 (range, 20–37; 13 female, 7 male) were recruited. Participants were recruited in the same way as in Experiment 1 but were told that the experiment was about the “perception of shapes.”

Stimuli and procedure

The string structures used were the same as those in Experiment 1. Order of stimulus presentation and timing parameters were also the same. For class A, eight blue rounded shapes were used; for class B, eight red angled shapes were used (Fig. 3). In auditory A^nB^n , the two syllable classes differed in speaker identity and vowel quality; in visual A^nB^n , the shapes differed in color and contour. Each shape was approximately 6×6 cm and was presented in series on a light gray background.

The procedure paralleled that used in Experiment 1, with the exception of the instructions. Only one set of instructions was used. The instructions for the training phase were the following: “Just relax and watch.” The instructions for the test phase were the following: “You will now see new shapes. Decide whether what you see fits to the examples you saw before. Decision time is important, so answer as quickly as possible.”

Results

The mean score for judgments based on A^nB^n was 80% ($SEM = 3.48$), with individual scores ranging from 50% to 98%. An independent samples *t* test showed no significant difference between male and female participants.

Fig. 3 Shapes used in visual A^nB^n learning. Shapes of class A were blue and round; shapes of class B red and angled

Group performance was analyzed by string type. Of the grammatical strings, 84.48% ($SEM = 3.78$) were accepted. As in [Experiment 1](#), the percentage of rejections varied across violation types (Fig. 4). The percentage of rejections was 65.42% ($SEM = 7.26$) for Type 1, 62.92% ($SEM = 7.39$) for Type 2, 71.67% ($SEM = 6.9$) for Type 3, 87.08% ($SEM = 4.78$) for Type 4, and 87.05% ($SEM = 3.55$) for Type 5. A repeated measures ANOVA for violation type showed a significant main effect, $F(4, 76) = 5.701$, $p < .001$. Violations of Type 2 were rejected significantly less often than violations of Type 4, $p = .001$, and Type 5, $p = .003$. Other significant differences did not survive Bonferroni correction (adjusted threshold: $p < .005$). Violations of Type 1 were rejected less often than violations of Type 4, $p = .006$, and Type 5, $p = .005$. Violations of Type 3 were rejected less than violations of Type 4, $p = .021$, and Type 5, $p = .023$.

The highest correlations were observed between the block consistency violations (Types 4 and 5), $r(18) = .753$, $p < .001$, and the counting violations (Types 1 and 2), $r(18) = .680$, $p < .001$. Counting violations and block consistency violations were each pooled for further analysis based on nonsignificant differences, high correlations, and structural similarities. The performance on counting violations was 64.17% ($SEM = 6.71$), and on block consistency violations, it was 87.06% ($SEM = 3.9$).

A repeated measures ANOVA showed significant differences between the grouped string types, $F(2, 38) = 5.628$, $p = .007$. Block consistency violations were rejected significantly more often than counting violations ($p = .001$; Bonferroni corrected significance threshold: $p < .017$). The difference between block consistency violation scores and

Type 3 violations was significant at the threshold for multiple comparison correction, $p = .019$. There was a positive correlation between counting and block consistency violation scores, $r(18) = .462$, $p = .04$. The correlation between Type 3 and block consistency violation scores was also positive, $r(18) = .502$, $p < .024$. There was no significant correlation between Type 3 and counting violation scores.

The results were compared with those of [Experiment 1](#). A series of independent sample t tests showed no significant differences in the percentages of rejections for any of the violation types (when compared individually or grouped) and no significant differences for the grammatical strings.

Template analysis

The results of the template analysis (using D scores) are presented in Fig. 5. The data suggest that of the 20 participants, 10 showed responses that were based on the acquisition of $A''B''$. One participant based responses on $(A''B'')+(B''A'')$, 1 on $(A^+B^+)+(B^+A^+)$, and 4 on A^+B^+ . The remaining four performances did not meet the criteria for being above chance.

As in [Experiment 1](#), there was no evidence that individual participant behavior on the first 60 trials was different from their behavior on the last 60 trials. For participants who performed above chance, the average D score for the best template was higher in the second half (86.07) than in the first half (76.29). A paired-sample t test showed this effect to be significant, $t(15) = 2.64$, $p = .019$.

Participant self-reports

Participant self-reports were grouped in the same fashion as in [Experiment 1](#) (Table 3). Participants whose performances were based on the acquisition of $A''B''$ all reported local dependencies in combination with counting. However, 2 participants who reported counting did not consistently show a counting-based performance. Participants who reported using no decision criteria, using intuition, or looking out for familiar units performed at chance.

Discussion

Overall, performance on the visual task was superior to that on the auditory task. However, an independent samples t test comparison of the number of correct judgments in each group showed that this difference did not reach significance, $t(38) = 0.996$, $p = .34$. Template analysis showed that only half of the participants consistently applied $A''B''$. It also revealed the same group-learning pattern that was

Fig. 4 Rejection patterns to the five violation string types in the visual condition. Bars represent percentages of correct rejections ($M \pm SEM$). Viol., violation type; cons., consistency

Fig. 5 Individual participant performance depicted as a function of best fit in the grammatical template analysis. See Table 1 for details of the eight possible grammatical templates. Single dots can correspond to several cases with identical values

observed in Experiment 1. Participants performing according to syntactic rules were able to report them after the experiment. It can be concluded that despite the differences in sensory modality, learning patterns in Experiments 1 and 2 were very similar.

General discussion

AGL is a paradigm used to investigate syntactic processing across different age groups, patient populations, and species. The aim of this study was to develop new methods for describing how participants process syntactically structured stimuli. In two experiments, the learning of the nonfinite grammar A^nB^n was explored first by looking at the group data, then by investigating individual behavioral

patterns. The group data suggested that experiments in both auditory and visual modalities replicated the results of previous studies (de Vries et al., 2008; Fitch & Hauser, 2004), since participants achieved similar accuracy levels in the test phase. Participant behavior was then investigated with a template analysis, which used different perspectives to look at the data for each individual. As was hypothesized, participants differed not only in how effectively they applied one set of decision criteria, but also substantially in their behavioral patterns. In both experiments, only about half of the individuals consistently rejected all types of violations of the target grammar. Others either rejected only a subset of violation types or performed at chance. Specifically, individual behavior could be attributed to the following templates: $(A^+B^+)(B^+A^+)$, A^+B^+ , $(A^nB^n)+(B^nA^n)$, or A^nB^n . There was no evidence that participants changed their behavior within the experiment. On the contrary, it appears that, with time, they became more consistent in applying their decision criteria. The present study showed no substantial differences between performances on visual stimuli and auditory stimuli with regard to scores and individual behavior. The comparison between specific and nonspecific instructions in Experiment 1 also yielded no differences.

The template analysis primarily describes group heterogeneity. However, it also allows insights into the processing of A^nB^n strings in general. In both experiments, participants who showed significant evidence of syntactic performance learned the following:

1. All items of one class preceded the items from the other class.

The detection of this pattern was implicit in all types of observed behavior— A^nB^n , A^+B^+ , $(A^nB^n)+(B^nA^n)$, and $(A^+B^+)(B^+A^+)$.

In addition to (1), participants detected none, one, or both of the following properties:

2. The first class was A [A^nB^n , A^+B^+].
3. The number of items in class A and B matched [A^nB^n , $(A^nB^n)+(B^nA^n)$].

Table 3 Self-reported decision criteria of individual participants in relation to the best grammatical template (numbers in parentheses represent chance-level performances)

Criteria reported after the task	Best Grammatical Template					
	A^nB^n	EVEN	$(A+B)^+B$	A^+B^+	$(A^+B^+)(B^+A^+)$	$(A^nB^n)+(B^nA^n)$
No criteria/intuition			(1)		(1)	
Familiar units		(1)	(1)			
Local				3		
Counting + local	10			1	1	1

On the basis of this description, it can be concluded that participants who managed to acquire A^nB^n did so by detecting all three structural properties (1–3), with (1) being the most salient.

The data show a clear relationship between the behavior noted under the template analysis and the reports given after the experiment. By the time the reports were given, participants were aware of their syntactic behavior (or lack thereof). However, it is not clear when the behavior becomes explicit. It is possible that learning takes place at an implicit level and becomes explicitly available at a later point in the experiment.

It would be mistaken to generally attribute the chance-level scores reported here to the lack of any pattern detection. The effects of *lexical behavior* were not examined—that is, participants making their decisions on the basis of the appearance or absence of specific syllables or shapes. Some participants reported such behavior (and were grouped under “familiar units” in the self-report analyses). Another explanation is that some participants applied idiosyncratic syntactic rules not contemplated in the template analysis. Ultimately, the analysis of AGL data can only rule out certain behaviors in the test phase. The templates chosen represent behaviors that were expected to be most likely. Given large enough sample sizes, regression models (Visser et al., 2009) can assist with the identification of behavioral templates. Since the template analysis was designed for small participant samples or even single cases, no such analysis was carried out.

Another issue concerning the template analysis is the possibility of *ties* between templates—that is, several templates having the same (or near-same) score. The likelihood of ties depends on two factors: how similar templates are to one another, and how many test trials are contained within the experiment. In the experiments reported here, the template A^+B^+ was very similar to the template $(A^+B^+)+(B^+A^+)$. The difference lay in the rejection of 12 violation strings—namely, those of Type 3 (B^nA^n). The D scores for the A^+B^+ and $(A^+B^+)+(B^+A^+)$ templates were, therefore, similar in each participant. In the data set reported here, exact ties did not occur. However, in some cases, the second highest D score was so close to the highest score that it was not possible to claim with confidence that the participant applied a specific set of decision criteria. However, the main results of these experiments are not affected by this issue. Participant behavior varied qualitatively, and some properties of grammatical strings were more salient than others (see above).

Although there were no observable differences in behavior between Experiments 1 and 2, this does not mean that the appearance of stimuli plays no role. Previously it has been reported that participants perform better when stimuli are produced by two speakers than when they are produced by one (Hochmann et al., 2008). It appears that performance is

dependent less on modality than on stimulus classes being distinguishable from each other. This suggests that experiments conducted with patient groups who have impairments in a particular modality, such as phonological processing, can use the results from a different modality for comparison.

The main advantage of the analysis of individual patterns is that it can reveal different kinds of behaviors among participants even if their total number of correct decisions in the test phase is similar. This might prove particularly useful in functional imaging studies that try to locate neural systems and networks responsible for specific syntactic processes. Neuroimaging techniques can be used to investigate brain activations in people with different behaviors. The methods reported here can also contribute to new studies investigating severely aphasic speakers. It has been observed that some people with agrammatism show atypical behavioral patterns when distinguishing (visual) A^nB^n strings from violations (Zimmerer & Varley, 2010). The template analysis can help describe these differences with greater precision.

A^nB^n strings have a relatively simple structure. A block of class A items is followed by a block of class B, and their numbers match. One can expect participant behavior to be at least as heterogeneous when more complex grammars are used. Methods like the ones introduced here, however, might turn heterogeneity from an analytical problem into a phenomenon that contributes to the understanding of syntactic processes.

Acknowledgements This research was supported by an ESRC Professorial Fellowship to R.A.V. We would like to thank the participants who took part in the experiments. We also thank Tom Fritzsch for his constructive comments on the manuscript, and Nigel Duffield for his helpful comments on this program of research. Further, we thank the three anonymous reviewers for their suggestions.

Appendix

Chance-level criteria for the template analysis.

To be categorized as above chance, the rejection of ungrammatical strings needed to be above chance, and the rejection of grammatical strings below chance.

Template	Chance level for ungrammatical strings rejected (%)	Chance level for grammatical strings rejected (%)
A^nB^n	35–65	35–65
EVEN	17–83	40–80
EQUAL	25–75	39–61
$A(A+B)^*$	25–75	39–61
$(B+A)^*B$	25–75	39–61
A^+B^+	31–69	38–62
$(A^+B^+)+(B^+A^+)$	25–75	39–61
$(A^nB^n)+(B^nA^n)$	33–67	37–63

References

- Bahlmann, J., Schubotz, R. I., & Friederici, A. D. (2008). Hierarchical artificial gprocessing engages Broca's area. *Neuroimage*, 42, 525–534.
- Caplan, D., Baker, C., & Dehaut, F. (1985). Syntactic determinants of sentence comprehension in aphasia. *Cognition*, 21, 117–175.
- Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23, 157–205.
- Cohen, M. C. (1991). *Introduction to computer theory* (Rev. ed.). New York: Wiley.
- de Vries, M. H., Monaghan, P., Knecht, S., & Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning: the learnability of embedded hierarchicals. *Cognition*, 107, 763–774.
- Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1984). A case of syntactical learning and judgment: how conscious and how abstract? *Journal of Experimental Psychology: General*, 113, 541–555.
- Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science*, 303, 377–380.
- Friederici, A. D. (2004). Processing local transitions versus long-distance syntactic hierarchies. *Trends in Cognitive Sciences*, 8, 245–247.
- Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature*, 440, 1204–1207.
- Goldberg, A. E. (2006). *Constructions at work. The nature of generalization in language*. Oxford: Oxford University Press.
- Hauser, M. D., Barner, D., & O'Donnell, T. (2007). Evolutionary linguistics: a new look at an old landscape. *Language Learning and Development*, 3, 101–132.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298, 1569–1579.
- Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition*, 78, B53.
- Herbranson, W. T., & Shimp, C. P. (2003). "Artificial grammar learning" in pigeons: a preliminary analysis. *Learning & Behavior*, 31, 98–106.
- Hochmann, J.-R., Azadpour, M., & Mehler, J. (2008). Do humans really learn *AnBn* artificial grammars from exemplars? *Cognitive Science*, 32, 1021–1036.
- Hoen, M., Golembiowski, M., Guyot, E., Deprez, V., Caplan, D., & Dominey, P. F. (2003). Training with cognitive sequences improves syntactic comprehension in agrammatic aphasics. *NeuroReport*, 14, 495–499.
- Jansen, B. R. J., Raijmakers, M. E. J., & Visser, I. (2007). Rule transition on the balance scale task: a case study in belief change. *Synthese*, 155, 211–236.
- Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, 283, 77–80.
- Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, 119, 264–275.
- Perruchet, P., & Rey, A. (2005). Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic Bulletin & Review*, 12, 307–313.
- Petersson, K. M., Forkstam, C., & Ingvar, M. (2004). Artificial syntactic violations activate Broca's region. *Cognitive Science*, 28, 383–407.
- Pinker, S., & Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, 95, 201–236.
- Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin*, 133, 227–244.
- Pothos, E. M., & Bailey, T. M. (2000). The role of similarity in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 847–862.
- Pothos, E. M., Chater, N., & Ziori, E. (2006). Does stimuli appearance affect learning? *The American Journal of Psychology*, 119, 277–301.
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behaviour*, 6, 855–863.
- Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: a reevaluation. *Journal of Experimental Psychology: General*, 125, 123–138.
- Rüsseler, J., Gerth, I., & Münte, T. (2006). Implicit learning is intact in adult developmental dyslexic readers: evidence from the serial reaction time task and artificial grammar learning. *Journal of Clinical and Experimental Neuropsychology*, 28, 808–827.
- Saffran, J. R. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928.
- Saffran, J. R., Hauser, M. D., Seibel, R., Kapfhamer, J., Tsao, F., & Cushman, F. (2008). Grammatical pattern learning by human infants and cotton-top tamarin monkeys. *Cognition*, 107, 479–500.
- Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27–52.
- Smith, J., Siegert, R. J., McDowall, J., & Abernethy, D. (2001). Preserved implicit learning on both the serial reaction time task and artificial grammar in patients with Parkinson's disease. *Brain and Cognition*, 45, 378–391.
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
- Uddén, J., Folia, V., Forkstam, C., Ingvar, M., Fernandez, G., Overeem, S., et al. (2008). The inferior frontal cortex in artificial syntax processing: an rTMS study. *Brain Research*, 1224, 69–78.
- van der Maas, H. L. J., & Straatemeier, M. (2008). How to detect cognitive strategies: commentary on 'Differentiation and integration: Guiding principles for analyzing cognitive change'. *Developmental Science*, 11, 449–453.
- Visser, I., Raijmakers, M. E. J., & Pothos, E. M. (2009). Individual strategies in artificial grammar learning. *The American Journal of Psychology*, 122, 293–397.
- Zimmerer, V. C., & Varley, R. A. (2010). Recursion in severe agrammatism. In H. van der Hulst (Ed.), *Recursion and human language* (pp. 393–406). Berlin: Mouton de Gruyter.