The Guessing Game: A Paradigm for Artificial Grammar Learning

Martin Redington

Dept. of Psychology
University of Edinburgh
7, George Square
Edinburgh, U.K.
EH8 9JZ

frankred@cogsci.ed.ac.uk

Nick Chater

Centre for Cognitive Science[1]
University of Edinburgh
1–4 Buccleugh Place
Edinburgh, U.K.
EH8 9LW

nicholas@cogsci.ed.ac.uk

Abstract

In a guessing game, Ss reconstruct a sequence by guessing each successive element of the sequence from a finite set of alternatives, receiving feedback after each guess. An upper bound on Ss knowledge of the sequence is given by $\hat{H}$, the estimated entropy of the numbers of guesses. The method provides a measure of learning independent of material type and distractors, and the resulting data set is very rich. Here, the method is applied to artificial grammar learning; Ss were exposed to strings from a finite state grammar and subsequently distinguished between strings that followed or violated the grammar reliably better than Ss who had not seen the learning strings (but who themselves performed at above chance levels). Ss knowledge of the strings, $\hat{H}$, reflected both grammaticality and exposure to learning strings, and was correlated with overall judgement performance. For non-grammatical strings, the strings that Ss knew most about were those they found most difficult to classify correctly. These results support the hypothesis that fragment knowledge plays an important part in artificial grammar learning, and we suggest that the guessing game paradigm is a useful tool for studies of learning and memory in general.

Introduction

In learning and memory research it is important to establish exactly what, and how much, Ss learn or recall in an experimental task. This paper describes how the "guessing game" paradigm (Shannon, 1951) can be useful for obtaining rich qualitative and quantitative information. We illustrate the approach with a guessing game variant on artificial grammar learning (AGL) tasks.

In a guessing game, Ss reconstruct a test sequence (or a set of sequences) by guessing each successive element of the sequence from a restricted set of alternatives, receiving feedback after each guess. This can be viewed as a collaboration between S and experimenter. The more information the S possesses about a sequence, the fewer guesses they will take when guessing which element comes next, and the less information (feedback) they will require from the experimenter.

Rather than measure how much Ss know about the items directly, we assess how much they do not know by estimating the entropy, H, of the sequence of numbers of guesses that they take (which is equivalent to that of the feedback they receive). This estimate, $\hat{H}$, is an upper bound, as Ss may know certain information, but be unable to use it in guessing, and because the guesses that even hypothetically ideal Ss take do not reveal the exact probability distribution of each item occurring next in the sequence, but only their order.

The originator of this method is Shannon (1951), who applied it to estimate the entropy of English, with Ss guessing each successive letter of an unseen text (an upper bound on Ss ignorance of the text is an upper bound on the entropy of the text, as the Ss uncertainty must be at least as great as the uncertainty, or entropy, of the text itself). Its application to psychological investigation was advocated by Attneave, who applied it to non-sequential, visual stimuli (Attneave, 1954).

In learning and memory experiments, the method can be used to assess the amount of information that Ss gain from the learning period, by comparing the "ignorance" of control Ss, with the "ignorance" of Ss in the learning condition. An important methodological advantage is that the guessing game allows learning performance to be measured in bits of information, thus allowing comparison across different sets and types of material. Another advantage is that it allows assessment of Ss' knowledge of each item independently of any foils or distractors. Furthermore, by requiring Ss to attempt to regenerate stimuli, a mass of low level detail about what Ss have learned, including memory for fragments and patterns rather than whole items, is provided.

The estimate, $\hat{H}$, of the amount of information in the feedback responses is given by:

$$\hat{H} = -\sum_{i=1,2,\ldots,n} \hat{p}_i \log_2(\hat{p}_i) \qquad (1)$$

where $\hat{p}_i$, the estimated probability that the S will re-

[1]Both authors are jointly affiliated.

quire i guesses to identify an element of the sequence, is derived from observed relative frequencies of i guesses being required[2].

Here, we apply this approach to the domain of AGL. In the classic AGL paradigm (Reber, 1967) Ss are asked to memorise a set of strings generated by a finite state grammar (see figure 1). Subsequently they are able to distinguish grammatical (G) strings from non-grammatical (NG) strings (that violate the grammar) at above chance and control levels. However, they cannot easily verbalise the knowledge that allows them to do this.

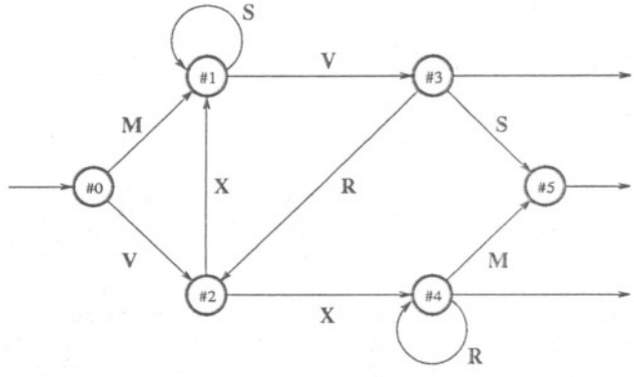

Figure 1: The finite state grammar; G strings are generated by following the paths starting at #0, and continuing until one of the 3 exiting paths is taken, with each path generating the letter that labels it.

This and related 'implicit' learning phenomena (*e.g.* serial reaction time tasks, Lewicki, Hill & Bizot, 1987) have been interpreted as evidence for a separate, unconscious learning mechanism, operating independently of awareness. This conclusion remains contentious, and a recent review concludes, premature (see Shanks & St. John, in press). Here, we are more interested in the nature and mechanisms of knowledge acquisition during AGL, and in why Ss should acquire some types of knowledge and not others.

Reber's claim for the AGL phenomenon (*e.g.* Reber, 1989) is that during exposure to the strings, Ss acquire a partial, but valid set of unconscious, abstract rules, which they use to distinguish G and NG strings. A contrasting position is that Ss acquire simple substring information, *e.g.* bigram frequency and position, permissible starting and ending letters, and use this to make their judgements. That Ss possess, and can verbalise, such knowledge is well established. Perruchet and Pacteau (1990) found that Ss could discriminate well between the bigrams (letter pairs) present in the training strings, and a new set of bigrams, and that strings composed of legal bigrams in an illegal order were more likely to be accepted as grammatical than strings of illegal bigrams. Dienes, Broadbent & Berry (1991) obtained similar results using a task which is effectively a restricted version of the present one—the sequential letter dependencies (SLD) task. Here, following the standard paradigm, Ss were presented with incomplete stems, *e.g.* MV..., and asked which letters could occur in the next position. Ss exhibited knowledge of both legal bigrams and their possible position. As well as being consciously accessible, Dienes *et al.* also found that this knowledge correlated with Ss grammaticality judgements, and both they and Perruchet and Pacteau found that Ss performance level on grammaticality judgements could be modelled using the bigram and position knowledge that they had expressed.

Ss clearly do exhibit some 'abstract' knowledge (knowledge which is independent of specific vocabulary items), as shown by transfer experiments (*e.g.* Matthews *et al.*, 1989) where in the test phase, the letters of the test strings are replaced with an entirely different set, but the deep structure of the items remains the same. Grammaticality judgement performance remains above chance and control levels.

Rather than focus on transfer here, we concentrate on validating the guessing game methodology, as a way of uncovering Ss' knowledge of substrings, letter transitions, and the relationship of these to Ss grammaticality judgements.

Method

Design: The experimental group was exposed to the set of learning strings before going on to perform the test phase. A control group performed only the test phase.

Materials: The grammar (see figure 1), learning strings, and test strings (see table 1) were those used by Dulany *et al.* (1984). The experiment was performed on Macintosh Colour Classic computers.

Subjects: The Ss from the experimental ($n = 36$) and control groups ($n = 10$) were unpaid undergraduate volunteers, aged between 19 and 42.

Procedure: Apart from the use of the guessing game paradigm, the experiment was closely modelled on that of Dulany *et al.* (1984). Prior to commencing the experiment, Ss were asked for their care and cooperation throughout the experiment, and this request was repeated after the test instructions. The following learning instructions were then displayed to the experimental Ss;

> This is a simple memory experiment. When you press the button labelled 'Start', you will see 20 items made of the letters M, R, S, V, and X. The items will run from three to six letters in length. Your task is to learn and remember as much as pos-

[2]This assumes that the sequence of feedback responses is independent. If there is sequential structure, then a more efficient coding of the sequence could be found by the S changing their guessing strategy, *e.g.* if the sequence 1, 2, 1, 2,... is reliably obtained, then a shorter code could be obtained if Ss simply made their second guess first on odd-numbered trials. If Ss use an optimal guessing strategy, there should be no sequential structure. The upshot of this is that $\hat{H}$ is an upper bound on the entropy of the stimulus. Additionally the use of observed probabilities can be a poor method with small samples, although here this effect is probably negligible.

Acquisition Grammatical	Grammatical	Test Non-Grammatical[2]
MSSSSV	VXSSSV	VXRRS
MSSVS	MSSSV	VXX
MSV[1]	MSSVRX	VXRVM
MSVRX	MVRXVS[1]	XVRXRR
MSVRXM	MSVRXV	XSSSSV[3]
MVRX	MSVRXR	MSVV
MVRXRR	MVRXM	MMVRX
MVRXSV	VXVRXR	MVRSR
MVRXV	MSSSVS	MSRVRX
MVRXVS[1]	VXRM	SSVS
VXM	MVS	MSSVSR
VXRR	MSVS	RVS
VXRRM	MSSV	MXVS
VXRRRR	MVRXR	VRRRM
VXSSVS	VXRRR	VVXRM
VXSVRX	VXSV	VXRS
VXSVS	VXR	MSRV
VXVRX[1]	VXVS[1]	VXMRXV
VXVRXV[1]	MSV[1]	MSM
VXVS[1]	VXRRRM	SXRRM
	VXSSV	MXVRXM
	VXV	MSVRSR
	VXVRX[1]	SVSSXV
	VXVRXV[1]	XRVXV
	MVRXRM	RRRXV

Table 1: The acquisition and test strings (after Dulany *et al.*, 1984). The letter T was replaced by S, in order to eliminate the familiar MTV item. [1]Strings present in both the acquisition and test sets. [2]Underlining indicates the point of grammatical violation. [3] The final V was accidentally omitted from this string, but this made little difference to the results.

sible about all 20 items. If you have any questions about the task, please ask the Experimenter now. Press 'Start' when you are ready to begin.

The learning strings were displayed for 10 minutes, in four left-justified columns of five strings each. The order of the strings was randomised, with the constraint that features of the grammar should not be made especially salient. Control Ss were simply informed that "another group of subjects saw items made up of the letters M, R, S, V, and X. We can't tell you what they were, but the items ran from three to six letters in length", before proceeding to the test phase. The test phase instructions ran as follows, with the variation for the control group in parentheses;

> The order of letters in the set you (they) saw was determined by a rather complex set of rules. The rules allow only certain letters to follow other letters. Now you will be presented with a set of test items. Your task is to guess the letters in the test strings. You can indicate your guesses by pressing the button corresponding to the letter that you think comes next. If your guess is incorrect, the button will disappear. If your guess is correct, the letter will appear on the screen, and you can proceed to the next letter. There is also a button labelled 'End' which you can press to indicate that the item is complete. The items are all between three and six letters long. You should try to make as few wrong guesses as possible.
>
> Once you have completed the item, two more buttons will appear, labelled 'Correct' and 'Incorrect'. If you think the item that you have just guessed follows the same rules as the original items, then press 'Correct'. If you think it violates those rules then press 'Incorrect'. Half of the test items follow the rules, and half violate them.

The test display initially showed a (blank) string display, centred on the screen, and below this, a row of guessing buttons, labelled from left to right, M, R, S, V, X and 'End'. Ss guessed by clicking on the appropriate button with the mouse. Following a wrong guess (*i.e.* not matching the next letter of the current item), the button disappeared. If their response was correct, then the letter was appended to the string display, and all the guessing buttons reappeared for the next letter to be guessed. During the guessing phase, the string display was right justified, giving no indication of the items' length. The 'End' button acted in an identical manner to the other guessing buttons; an 'End' guess was correct if the item was otherwise complete (all its letters had been guessed), and incorrect otherwise. Following a correct 'End' guess, the guessing buttons disappeared, the string display was centred, and two buttons labelled 'Correct' and 'Incorrect' appeared to the right of the string display. After the S responded by clicking one of these, the test display was reset for the next item.

The 25 G and 25 NG test items were each repeated once, and the resulting 100 test items presented in random order. Due to the large number of test items (and consequently guesses to be made), the test took up to an hour to complete. Ss were permitted a short (5–10 minute) break during the test phase if they so wished.

Results

Judgements: Analysis of Ss grammaticality judgements revealed that, in terms of mean proportions correct, relative proportions of error pairs, and rankings of string difficulty, they mirrored those of earlier, comparable studies (Reber & Allen, 1978; Dulany *et al.*, 1984; Dienes, Broadbent & Berry 1991) suggesting that the use of the guessing game did not contaminate performance by introducing influences absent in the original paradigm.

The experimental Ss showed no overall response bias, although a few Ss were notably biased, and there was a significant rank correlation between bias (the ratio of G and NG responses) and the total proportion of correct responses (Spearman's $\rho = -.39, p < 0.01$). Control Ss were significantly biased towards G responses (means of 0.71 G responses to 0.29 NG $t(9) = 4.05, p < 0.01$), and showed a similar correlation ($\rho = -.61, p < 0.05$).

The proportion of strings correctly classified by the experimental Ss significantly exceeded that of the con-

Group:	Experimental (n = 36)		Control (n = 10)	
String Responses	Mean	S.D.	Mean	S.D
correct-correct	0.46	0.11	0.41	0.07
error-error	0.22	0.08	0.28	0.12
error-correct	0.16	0.07	0.17	0.09
correct-error	0.16	0.05	0.14	0.09
mean mixed	0.16	0.06	0.15	0.09

Table 2: The mean proportion of strings that Ss judged correctly or erroneously across both presentations of the string. As in previous studies, the error-error category is significantly higher than either of the mixed cases.

trols (means = 0.619 and 0.566 respectively, $t(44) = 1.95, p < 0.05$). For experimental Ss, this value was slightly lower than the 0.695 of comparable Ss from Dulany *et al.* (1984), and the range of values, 0.49–0.84, was wider than the 0.63–0.70 found by Dulany *et al.* Thus some Ss learnt little or nothing from their exposure to the learning set, whilst others performed at the same level as Reber and Allen's (1978) 'specially selected advanced ...students'.

Interestingly, control Ss performance was significantly greater than the 0.5 expected by chance ($t(9) = 3.78, p < 0.005$). This surprising effect was also observed by Dulany *et al.* (1984), whose control Ss got 0.555 of judgements correct. However, neither group showed evidence of significant improvement between the first and second presentation of each test item (experimental Ss means were 0.621 & 0.617 ($t(35) = 0.39$, control Ss; 0.552 & 0.580, $t(9) = 1.29$).

Both groups showed an inflated error-effect as found by Dulany *et al.* (1984) and Dienes (1992), as shown in table 2. The number of error-error judgement pairs was greater than both the error-correct (experimental Ss; $t(35) = 3.28, p < 0.005$, control Ss; $t(9) = 1.72, p = 0.059$) and correct-error (experimental Ss; $t(35) = 3.25, p < 0.005$, control Ss; $t(9) = 2.20, p < 0.05$) judgement pairs.

Finally, Dienes (1992) found consistent rankings of string difficulty (based on the number of Ss who correctly classified each string). The rank correlation between our experimental Ss rankings of string difficulty, and Dienes' TOTAL rankings (based on the combined data from Dulany *et al.* (1984), and Dienes, Broadbent & Berry (1991)), was not significant for G items ($\rho = 0.28, p > 0.05$), but was comparable for NG items ($\rho = 0.59, p < 0.005$). However, if the non-learners (the 6 Ss whose judgement performance was less than 0.538, 1 standard deviation below the mean) were excluded, the correlations rose to $\rho = 0.43, p < 0.025$ for G items and $\rho = 0.63, p < 0.005$ for NG items. For control Ss, only the correlations for NG items were significant (for G items, $\rho = 0.09$ and for NG $\rho = 0.45, p < 0.025$).

Guesses: The guessing game provides a very rich set of data, and we present only a tiny subset of the many possible analyses here. Both $\hat{H}$, calculated as in 1, and the mean number of guesses per letter (GPL) are reported. As $\hat{H}$ is equivalent to the number of yes/no questions required to identify an item (where each question reduces the uncertainty by half), it is highly correlated (Pearson's correlations of 0.95 or more) with GPL.

A 2 × 2 × 2 (group × grammaticality × presentation) split-plot ANOVA, on Ss mean $\hat{H}$, revealed significant main effects for group ($F(1, 44) = 5.56, p < 0.025$), grammaticality ($F(1, 44) = 40.63, p < 0.001$), and presentation ($F(1, 44) = 20.16, p < 0.001$). Thus all Ss exhibited more knowledge about G items than NG items, and experimental Ss (who took a mean 2.51 and 2.78 GPL for G and NG items) exhibited more knowledge than control Ss (who took 2.75 and 3.04 GPL). The improvement across presentations, while significant, was minor (over both groups, and all strings, Ss took a mean 0.11 fewer guesses on the second presentation of each item. The only significant interaction effect was between grammaticality and presentation ($F(1, 44) = 8.13, p < 0.01$), Ss improved more across presentations of G items than NG items. However, this difference was also small, and significant only for $\hat{H}$, not GPL.

For experimental Ss, a significant correlation was found between the proportion of strings that Ss classified correctly, and the knowledge they exhibited in their guesses, $\hat{H}$, measured over all the strings ($\rho = -.54, p < 0.05$)—Ss who were good at classifying also tended to be good at guessing. This is much higher than the (0.29) correlation found by Dienes, Broadbent & Berry (1991) using the SLD task. Again, control Ss exhibited a similar relationship ($\rho = -.62, p = 0.05$).

For NG items, both experimental and control Ss showed a strong correlation between the rankings of string difficulty for the group, and Ss knowledge of each string, $\hat{H}$, averaged over the Ss in each group ($\rho = .79, p < 0.005$ and $\rho = .55, p < 0.005$ respectively). Thus the strings that Ss knew most about (took fewest guesses on) were also the ones that they were most likely to misclassify. For G items this pattern was not so clear, the corresponding correlations were $\rho = -.22, ns$ for experimental Ss, remaining unreliable when non-learner were eliminated (as before) and $\rho = -.36, p < 0.05$ for controls. The direction of these correlations suggests that the G strings that Ss knew least about were most likely to be misclassified.

Discussion

The guessing game results support the findings of Dulany *et al.* (1991) and Perruchet and Pacteau (1990) that Ss exposed to the learning strings can exhibit considerable knowledge of the letter transitions permitted by the grammar (hence their higher knowledge, $\hat{H}$, of the G strings).

The strong correlation between Ss guessing and judgement performance can also be seen as supporting the notion that the knowledge expressed in the guessing game also mediates grammaticality judgements. However, caution should be used, as it is likely that other factors (such as different Ss attention and motivation) affect both scores, and account for at least part of this relationship.

A more interesting finding is that for NG strings, the items that Ss take least guesses on (know most about), are the ones they find hardest to classify correctly. This is consistent with Ss using their transition knowledge and a familiarity heuristic in making their judgements—strings which contain unfamiliar transitions (which take Ss more guesses as they guess the expected items first, and then try the less likely possibilities) are classified as NG. However, the failure to find this effect with G strings is disappointing. This may be due to the fact that our Ss performance was fairly noisy, but in both this and other unpublished studies we have found that NG strings' rankings of difficulty are more consistent than those of G strings.

Our most peculiar finding was the above chance judgement performance of the control group (as found by Dulany *et al.*, 1984). We suggest that both the guessing game and Dulany *et al.*'s underlining task (where Ss underline that part of each string which in their view makes it G or NG) caused Ss to attend to the structure of the strings more than the standard paradigm. Many of our control Ss reported informally that they believed that only M and V could legally start a string (which is true for all G items, and 68% of NG items). Of the 8 strings that violate this rule, 6 were amongst the 10 easiest NG strings for control Ss. Thus by attending carefully to the structure of the test strings, even control Ss appear to be able to learn something about the structure of the grammar, and this may largely account for their performance.

The results presented here barely summarise the rich set of guessing data that this paradigm provides, and further analyses will hopefully reveal the degree to which Ss learn the different possible transitions, and why some transitions are easier to learn than others (as appears to be the case). Additionally, we have constructed computational models of the AGL process, and tested them against the guessing game data. Our present models (which learn by reducing the redundancy in their representation of the learning strings) easily match Ss judgement level of performance, but not rankings of difficulty, or proportion of error-pairs. The guessing game provides fine-grained data against which we can assess the success of such models.

Whilst we believe that the fragment knowledge demonstrated here and elsewhere plays an important part in AGL, transfer experiments suggest it is not the whole story, and we are currently performing transfer experiments using the guessing game paradigm. Data from current and future guessing game studies will hopefully shed light on the nature of the learning processes involved, and the knowledge that Ss possess. Finally, we believe that the guessing game's potential has been long overlooked, and that it deserves attention as a tool not only for the study of AGL, but for learning and memory research in general.

Acknowledgements

This research was supported in part by ESRC grant No. R00429234268, to the first author. Data collection was performed by the first author, Rhys Robins, Caroline Baillie, Caroline Smith, Karen Cockburn, Sean P. Cumming, Joel Levy, Katherine Sutton and Kirsty Reid. Thanks to Zoltan Dienes who kindly provided the the TOTAL rankings data, and Don Dulany, who advised on the test conditions and materials.

References

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, **61**, 183–193.

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science*, **16**, 41–79.

Dienes, Z., Broadbent, D.E., & Berry, D.C. (1991). Implicit and explicit knowledge bases in artificial grammar learning. *Journal of experimental psychology: Learning, Memory, & Cognition*, **17**, 875–887.

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.

Lewicki, P., Hill, T. & Bizot, E. (1988). Acquisition of procedural knowledge about a pattern of stimuli that cannot be articulated. *Cognitive Psychology*, **20**, 24–37.

Mathews, R.C., Buss, R.R., Stanley, W.B., Blanchard-Fields, F., Cho, J.R. & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of experimental psychology: Learning, Memory, & Cognition*, **15**, 1083–1100.

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

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, **5**, 855–863.

Reber, A.S., & Allen, R. (1978). Analogy and abstraction strategies in synthetic grammar learning: A functional interpretation. *Cognition*, **6**, 189–221.

Shanks, D.R., & St. John, M.F. (in press). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences.*

Shannon, C.E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, **30**, 50–64.