Chunking or Not Chunking? How Do We Find Words in Artificial Language Learning?

Ana Franco (afranco@ulb.ac.be)

Arnaud Destrebecqz (adestre@ulb.ac.be)

Cognition, Consciousness and Computation Group
Université libre de Bruxelles, 50 ave. F.-D. Roosevelt, B1050 BELGIUM

Abstract

What is the nature of the representations acquired in implicit statistical learning? Recent results in the field of language learning have shown that adults and infants are able to find the words of an artificial language when exposed to a continuous auditory sequence consisting in a random ordering of these words. Such performance can only be based on processing the transitional probabilities between sequence elements. Two different kinds of mechanisms may account for these data: Participants either parse the sequence into smaller chunks corresponding to the words of the artificial language, or they become progressively sensitive to the actual values of the transitional probabilities. The two accounts are difficult to differentiate because they tend to make similar predictions in similar experimental settings. In this study, we present two experiments aimed at disentangling these two theories. In these experiments, participants had to learn two sets of pseudo-linguistic regularities (L1 and L2) presented in the context of a Serial Reaction Time (SRT) task. L1 and L2 were either unrelated, or the intra-words transitions of L1 became the inter-words transitions of L2. The two models make opposite predictions in these two situations. Our results indicate that the nature of the representations depends on the learning conditions. When cues are presented to facilitate parsing of the sequence, participants learned the words of the artificial language. However, when no cues were provided, their performance was strongly influenced by the actual values of the transitional probabilities.

Keywords: implicit statistical learning; SRN; chunking; serial reaction time task

Introduction

A central issue in implicit learning research concerns the nature of the acquired knowledge. Does it reflect the abstract rules on which the training material is based or the surface features of the material, such as the frequencies of individual elements or chunks? According to some theorists, cognition can be viewed as rule-based symbol manipulation (Pinker & Price, 1988). From this perspective, learning would consist in the formation of new abstract, algebra-like rules. According to another theoretical position, information processing is essentially based on associative processes. In this view, learning would not depend on rule acquisition but on mechanisms capable of extracting the statistical regularities present in the environment (e.g., Elman, 1990).

Over the last few years, a series of experimental results have provided new insights into the question of the nature of the representations involved in implicit learning. Research on language acquisition has shown that 8-months old infants are sensitive to statistical information (Jusczyk et al., 1994; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999) and capable of learning distributional relationships between linguistic units (Gomez & Gerken, 1999; Jusczyk, Houston, & Newsome, 1999; Saffran, Aslin, & Newport, 1996; Perruchet & Desaulty, 2008) presented in the continuous speech stream formed by an artificial language.

Other studies have indicated that adults are also capable of extracting statistical regularities, and that these mechanisms are not restricted to linguistic material but also apply to auditory non-linguistic stimuli (Saffran, Johnson, Aslin, & Newport, 1999) or to visual stimuli (Fiser & Aslin, 2002).

In the same way, implicit sequence learning studies have indicated that human learners are good at detecting the statistical regularities present in a serial reaction time (SRT) task. Altogether, these data suggest that statistical learning depends on associative learning mechanisms rather than on the existence of a "rule abstractor device" (Perruchet, Tyler, Galland, & Peereman, 2004). However, different models have been proposed to account for the data. According to the Simple Recurrent Network model (Elman, 1990; Cleeremans, & McClelland, 1991; Cleeremans, 1993), learning is based on the development of associations between the temporal context in which the successive elements occur and possible successors. Over training, the network learns to provide the best prediction of the next target in a given context, based on the transitional probabilities between the different sequence elements. On the other hand, chunking models, such as PARSER, consider learning as an attention-based parsing process that results in the formation of distinctive, unitary, rigid representations or chunks (Perruchet & Vinter, 1998). Thus, both models are based on processing statistical regularities, but only PARSER leads to the formation of "word-like" units.

Although the representations assumed by these two classes of models are quite different, contrasting their assumptions is made difficult by the fact that they tend to make similar experimental predictions. For instance, in a typical artificial language learning experiment, participants are exposed to a continuous stream of plurisyllabic non-words (e.g., BATUBI, DUTABA…) presented in a random order, such that transitional probabilities between syllables are stronger intra-word

than between words. As the representations that emerge in either model reflect the strength of the associations between elements, both predict faster processing for intra-words than for inter-words transitions as well as successful recognition of the artificial language words.

In order to contrast the predictions of these two models, we used a Serial Reaction Time (SRT) task in which participants had to learn two different artificial languages presented successively. In one (control) condition, the two languages were not related to each other, but in the other (experimental) condition, the intra-words transitions of the first language (L1) became inter-words transitions in the second language (L2). Two different hypotheses can be formulated. On the one hand, if learning depends on chunk formation, the probability that one element will follow another is 100% within-words and 0% between-words. In order to learn L2 words, participants must first break the chunks formed during training on L1 and then form the new L2 chunks. This task should be easier in the control than in the experimental group since, in the former case, L1 transitions are no longer presented during L2. L1 chunks will then progressively decay and be replaced by L2 chunks. By contrast, in the experimental condition, L1 transitions are still presented, although less frequently, between L2 words presentation. As a result, L1 chunks continue to be reinforced during L2 presentation. It will then take more processing time in order to replace L1 chunks by L2 chunks in the experimental condition. One might therefore expect better recognition of L2 "words" in the control than in the experimental condition

On the other hand, if learning strictly reflects transitional probabilities, the probability that one element will follow another is 100% within-words and 33% between-words —since there are 4 different words and no repetitions. Thus, when switching from L1 to L2, the SRN must either develop new associations between elements (in the control condition) or merely "tune" the strength of the association between sequence elements (in the experimental condition). One might therefore expect better recognition of L2 "words" in the experimental than in the control condition[1].

Experiment 1

Participants

Twelve undergraduate students of the Université Libre de Bruxelles took part in the experiment in exchange for course credit. All reported normal or corrected-to-normal vision.

Apparatus and display

The experiment was run on a Mac mini computer equipped with a tactile monitor. The display consisted of twelve invisible dots arranged in a square on the computer's screen. Each dot represented a possible position of the visual moving target.

The stimulus was a small red circle 0.65 cm in diameter that appeared on a gray background, centered 0.10 cm below one of the twelve invisible dots separated by 2.20 cm.

Procedure

The experiment consisted of 9 training blocks during which participants were exposed to two different language-like sequences in a serial reaction time task. In the three first training blocks, they were exposed to a first language (L1) composed by four two-location "words" or sequences. Each word was presented 200 times, for a total of 1600 trials. In the six subsequent blocks, participants were exposed to a second language (L2) composed by four three-location words presented 250 times each, for a total of 3000 trials. On each trial, a stimulus appeared at one of the possible twelve positions. Participants were instructed to press the location of the target as fast as possible with the ad hoc pen. The target was removed as soon as had been pressed, and the next stimulus appeared either after a 250 msec response-stimulus interval (RSI) for intra-words transitions or a 750 msec RSI for inter-word transitions. Participants were not informed that the sequence of locations corresponded to the succession, in a random order, of the four "words" of the artificial languages. They were allowed to take short rest breaks between any two blocks.

Participants were randomly assigned to two conditions. In the control condition L1 and L2 were unrelated and in the experimental condition the intra-word transitions of L1 became inter-words transitions in L2. Thus, whereas L1 differed between control and experimental conditions, L2 was the same in both conditions.

All participants were subsequently asked to perform a recognition task in which they were required to decide if they had been exposed to each sequence during the training phase or not. Three types of sequences were presented: 8 "words" from L2 (each sequence presented twice), 4 "part-words" (sequences spanning L2 word boundaries) and 4 "non-words", which corresponded to visual sequences which had never been presented during L2 training.

Stimulus material

The display consisted of twelve invisible dots arranged in a square on the computer's screen. Each dot represented a possible position to the visual moving target.

The stimulus set consisted of sequences of word-like units in which the visual target could take 12 possible positions (numbered to 1 to 12). In the control condition, L1 consisted in four two-location words: 3-1, 6-4, 9-7 and 12-10. In the experimental condition, the words were: 3-4, 6-7, 9-10 and 12-1. In both conditions, L2

[1] This methodology is based on an original idea by Ronald Peereman and Pierre Perruchet.

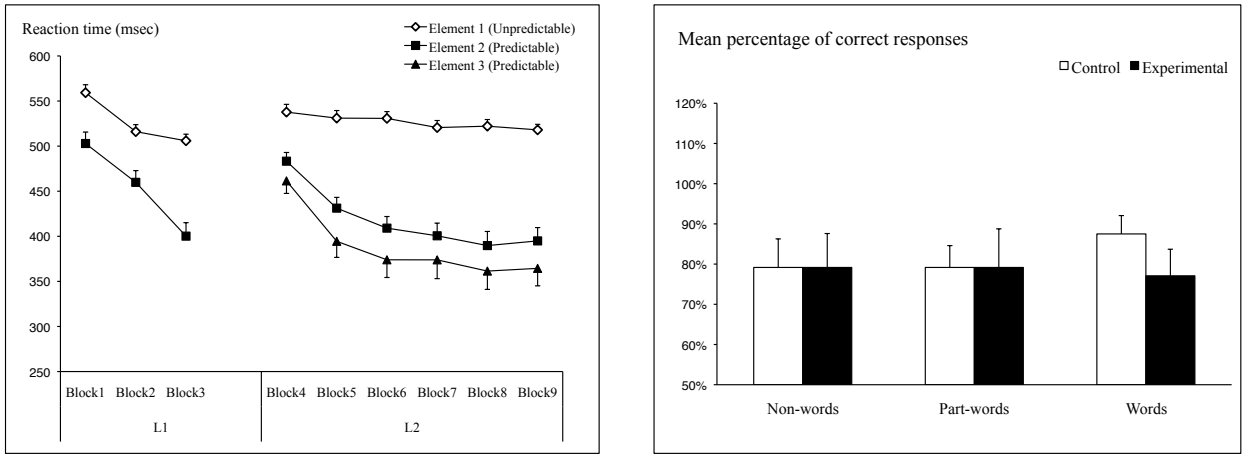

Figure 1. The figure shows mean RTs obtained for unpredictable (element 1) and predictable elements (elements 2 and 3) during L1 and L2 blocks. RTs are averaged over experimental and control conditions (left panel). Mean percentage of correct responses during the recognition task for words, non-words and part-words in the control and experimental conditions are displayed on the right panel. 50% is chance level.

consisted in four three-location words: 1-2-3, 4-5-6, 7-8-9 and 10-11-12. Stimuli were presented in a pseudo-random manner: A word was never followed by itself.

A different mapping between the 12 sequence elements and the 12 screen locations was used for each participant.

Results

RT task

To assess whether participants were able to learn L1 and L2, we examined separately mean reaction times (RTs) for the first three blocks (L1) and the next six blocks (L2) in the control and experimental conditions. Recall that the stimulus material was such that the first element of each word-like unit was unpredictable, whereas the second (and third in L2) were completely predictable. Figure 1 (left panel) shows the average reaction times obtained over the entire experiment, plotted separately for each element of the word-like sequences. Given that participants performed similarly in the control and experimental conditions ($F(1,10)=2.113$, $p>.1$ for L1 and $F(1,10)=.481$, $p>.5$ for L2), we pooled them together. The figure makes it clear that participants' responses are strongly influenced by the serial position within each sequence: RTs decreased more and were faster for predictable elements than for unpredictable elements. Two two-way analyses of variance (ANOVA) conducted on mean reaction times confirmed these impressions. First we examined the first three blocks (L1) by using an ANOVA with block [3 levels] and element [2 levels – predictable and unpredictable] as repeated measures factors. This analysis revealed a significant main effect of Block, $F(2,10)=56.007$, $p<.0001$, and Element, $F(1,10)=15.431$, $p<.005$. The interaction also reached significance, $F(2,10) = 6.630$, $p< .01$. Second we examined the next six blocks (L2) by using an ANOVA with Block [6 levels] and Element [3 levels] as repeated measures factors. A significant main effect of Block was found, $F(5,50)= 15.113$, $p<.0001$. The analysis also revealed a significant main effect of Element, $F(2,20)= 25.141$, $p< .0001$. The interaction also reached significance, $F(10,100) = 6.220$, $p<.0001$.

Recognition task

Figure 1 (right panel) shows recognition performance for the three types of test sequences plotted separately for control and experimental conditions. The figure indicates that the participants recognized L2 words, non-words and part-words in the two conditions. These impressions are confirmed by a series of one-tailed t-tests (see Table 1).

Table 1: *t* values comparing recognition scores to chance level in control and experimental conditions for the three types of test sequences. * indicates that the test reached significance (one-tailed, $p < .05$).

	Words	Non-words	Part-words
Control	5.82*	2.91*	3.79*
Experimental	2.89*	2.44*	2.15*

More importantly, performance was reliably better for L2 words in the control condition as compared to the experimental condition, one-tailed $t(47) = 1.70$, $p < .05$. As clearly illustrated on Figure 1 (right panel), all the other comparisons were not significant.

Discussion

SRT results indicate that participants learned the first and second "languages" in both the experimental and control conditions. The recognition results showed that participants were able to discriminate the word-like units of the second language. Importantly, performance was improved in the control condition as compared to the

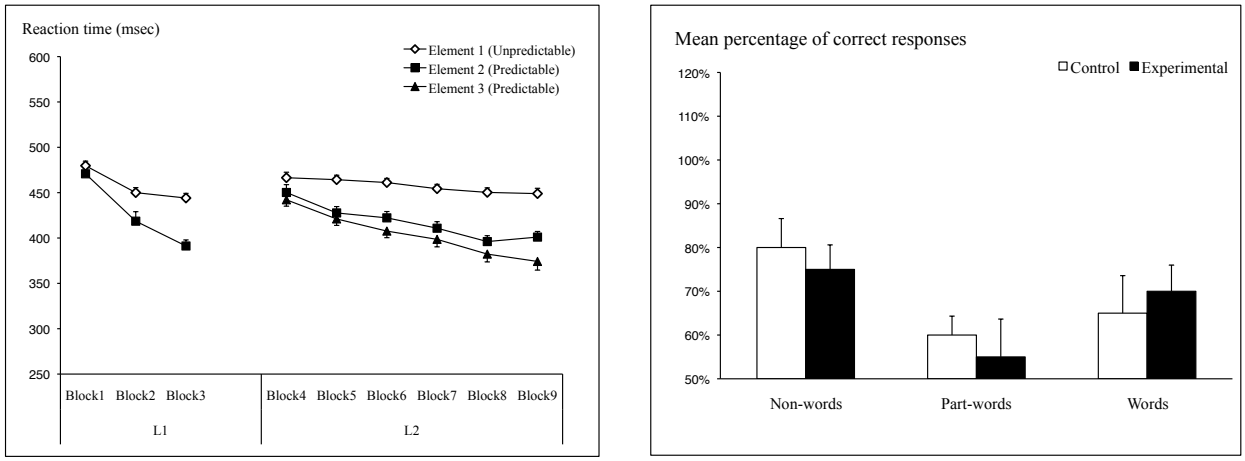

Figure 2. The figure shows mean RTs obtained for unpredictable (element 1) and predictable elements (elements 2 and 3) during L1 and L2 blocks. RTs are average over experimental and control conditions (left panel). Mean percentage of correct responses during the recognition task for words, non-words and part-words in the control and experimental conditions are displayed on the right panel. 50% is chance level.

experimental condition, i.e. when the two language-like sequences did not share any transitions between elements.

Taken together, these results are in line with the notion that participants learn the word-like sequences based on parsing mechanisms.

Recall that, in the experimental condition, L1 transitions (e.g., 3-4) were still presented between words during L2 presentation (e.g., between 1-2-3 and 4-5-6). As a result, L1 chunks continue to be reinforced during L2 presentation. As a result, a chunking model such as PARSER would predict better L2 recognition in the control than in the experimental condition. The observation that non-words and part-words rejection did not differ between these two conditions also fits with the prediction of a chunking model. The representational units that result from learning in such a model do not reflect the actual transitional probabilities present in the training sequence. The probability to erroneously consider a test sequence as a word of L2 should not be higher for part-word than for non-words even though the transitional probabilities between elements are higher in the former cases.

In Experiment 1, however, word-like sequences were clearly identified by the use of a larger RSI for inter-words than for intra-word transitions. Therefore, it remains possible that our results depend on this particular presentation mode. To address this possibility, we conducted a second experiment in which the RSI was set to a constant value.

Experiment 2

Participants

Ten undergraduate students of the Université Libre de Bruxelles took part in the experiment in exchange for course credit. All reported normal or corrected-to-normal vision.

Apparatus and Display

The apparatus and display were identical to those used in Experiment 1.

Procedure

The procedure was identical to the one used in Experiment 1 except that the RSI was fixed at 250 msec for intra-word transitions and inter-word transitions.

Stimulus material

The stimuli were identical to those used in Experiment 1.

Results

Reaction Time

Figure 2 (left panel) shows the average reaction times obtained over the entire experiment, plotted separately for each element of the word-like sequences. As in Experiment 1, we pooled control and experimental conditions together since there was no difference in performance between both conditions ($F(1,8)= 1.114$, $p >.1$ for L1 and $F(1,8)= .042$, $p >.5$ for L2). The figure clearly indicates that RTs are strongly influenced by the position: RTs decreased more and were faster for predictable elements than for unpredictable elements.

Two two-way ANOVA conducted on mean RTs confirmed these impressions. First we examined the first three blocks (L1) by using an ANOVA with Block [3 levels] and Element [2 levels – predictable and unpredictable] as repeated measures factors. This analysis revealed a significant main effect of Block, $F(2,16)=37.227$, $p<.0001$ and of Element, $F(1, 8) = 9.720$, $p<.05$. The interaction also reached significance, $F(2, 16) = 7.337$, $p<.005$. Second we examined the next six blocks (L2) by using an ANOVA with Block [6 levels] and Element [3 levels] as repeated measures factors. We found a significant main effect of Block,

F(5, 40) = 9.657, p<.005. The analysis also revealed a main effect of Element, F(2, 16) = 8.404, p<.005. The interaction also reached significance, F(10, 80) = 6.914, p<.0001.

Recognition task

Correct recognitions are plotted in Figure 2 (right panel). As indicated in Table 2, participants were able to correctly reject non-words in both conditions. They did not, however, correctly reject part-words. Concerning L2 words, experimental participants recognize them above chance but this was not the case in control participants.

Table 2: *t* values comparing recognition scores to chance level in control and experimental conditions for the three types of test sequences. * indicates that the test reached significance (one-tailed, p < .05).

	Words	Non-words	Part-words
Control	1,24	3.21*	1,63
Experimental	2.36*	3.17*	0,41

Overall, performance did not significantly differ between control and experimental conditions (all *p*s >. 05). Therefore, we pooled control and experimental conditions together and compared performance for non-words, part-words and L2 words. This analysis revealed a significant difference between non-words and part-words, paired t(78)= 1.574, p <.05: non-words rejection was better than part-words rejection (see Figure 2, right panel). The other comparisons failed to reach significance.

Discussion

In Experiment 2, L1 and L2 were presented using a constant RSI. However, relying exclusively on the only available cue, i.e. the statistical regularities, participants learned the first and second languages. Indeed, throughout training, mean RTs decreased more for predictable than for unpredictable elements. Moreover, participants recognized L2 words, at least in the experimental condition and correctly rejected non-words. Interestingly, in both experimental and control conditions, participants performed better in rejecting non-words than part-words, which were not correctly rejected.

According to PARSER, performance should be the same for non-words and part-words. If participants formed L2 chunks during training, it should be as easy to reject non-words than part-words as these sequences do not match the units formed during training. On the contrary, the SRN predicts that participants should recognize L2 words, which correspond to high transitional probabilities and reject non-words, which correspond to low transitional probabilities. However, as part-words involved high transitional probabilities, the SRN may have more difficulties in rejecting them. Experiments 2 results fit nicely with the SRN predictions, suggesting that participants are indeed sensitive to the actual values of the transitional probabilities between sequence elements. When considering Experiments 1 and 2 together, our results suggest that the values of transitional probabilities influence performance when no temporal cues guide the chunking process.

General Discussion

In this paper, we aimed at clarifying the nature of the representations involved in implicit and statistical learning. The question is to assess whether participants form chunks of the training material or merely develop a sensitivity to the transitional probabilities present in the training sequence. We showed that, in the context of a visuo-motor reaction time task, participants learn the statistical regularities present in a random succession of word-like sequences of visual targets. They are able to learn two different languages (L1 and L2) presented successively. Moreover, they are also able to recognize the word-like units of L2 in a subsequent recognition task. When word-like units are clearly separated from each other, recognition performance is improved in a control condition in which L1 and L2 do not share any pairwise transitions between language elements. These results are in line with the notion that word-like, rigid, disjunctive units are developed during learning. However, chunk formation seems not to be automatic. When the word-like units are not clearly marked – i.e. when they are presented in a continuous stream without any temporal cue to guide the chunking process – recognition performance is more influenced by the actual values of the transitional probabilities between sequence elements. This is reflected in Experiment 2 by better rejection of non-words than part-words in the recognition task.

Another potential explanation for this result could be that participants did form chunks in Experiment 2 but that they did not correspond to the actual L2 words. It is possible that participants indeed parsed the continuous sequence of visual stimuli into smaller chunks but that these chunks did not respect the actual boundaries between words. They may have focused, for instance, on particularly salient transitions (for example between elements that were spatially close to each other or between alternating locations) and end up with larger, smaller or different chunks than those corresponding to the words of the artificial language. In other words, if chunking is not directly induced by the presentation mode, attentional factors may also influence chunk formation. As a consequence, the actual chunks may differ from one participant to another and may not strictly reflect the transitional probabilities between the different sequence elements. This may, of course, influence recognition performance.

Both the SRN and PARSER implement elementary associative learning mechanisms such that, in both cases, the system tends to associate elements that occur often in succession. As a consequence, even if the chunks resulting from training do not correspond to the actual

words of the artificial language, there is a good chance that they involve highly frequent transitions. Participants may therefore tend to erroneously consider these part-words as words of the artificial language because they involve such high-frequency transitions.

In summary, this study shows that when units are marked, the chunking models provide reliable assumptions concerning the nature of the representations developed during learning. However, when no cues are provided in order to guide the chunking processes, performance reflects sensitivity to the strength of the transitional probabilities and seems also to depend on attentional factors. Further modeling studies are needed in order to test the ability of the SRN and PARSER models to account for the experimental results described in our study.

Acknowledgments

This research was supported by a FNRS grant number 1.5.057.07F to AD. AF is supported by a FNR governemental grant (Luxembourg).

References

Cleeremans, A. (1993). Mechanisms of implicit learning : Connectionist models of sequence processing. Cambridge: MIT Press.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, 120, 235-253.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24), 15822-15826.

Gomez, R. L. & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, 70, 109–135.

Jusczyk, P. W., Houston, D. M. & Newsome, M. The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39, 159–207 (1999).

Jusczyk, P. W., Luce, P., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33, 630–645.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory & Language*, 39, 246-263.

Perruchet, P., Tyler, M., Galland, N., & Peereman, R. (2004). Learning nonadjacent dependencies: No need for algebraic-like computations. *Journal of Experimental Psychology: General*, 133(4), 573-583.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? Memory & Cognition, 36, 1299-1305.

Pinker, S., & Price, A. (1988). On language and connectionism: Analysis of parallel distributed processing model of language acquisition. *Cognition*, 28, 73-193.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926-1928.

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(1), 27-52.