PERGAMON

Journal of Neurolinguistics 17 (2004) 97–119

Journal of NEUROLINGUISTICS

www.elsevier.com/locate/jneuroling

The exploitation of distributional information in syllable processing

Pierre Perruchet*[,1], Ronald Peereman[1]

LEAD/CNRS, Université de Bourgogne, Pole AAFE, Esplanade Erasme, 21000 Dijon, France

Received 27 February 2003; received in revised form 1 July 2003; accepted 11 July 2003

Abstract

There is now growing evidence that people are sensitive to the statistical regularities embedded into linguistic utterances, but the exact nature of the distributional information to which human performance is sensitive is an issue that has been surprisingly neglected as yet. In order to address this issue, we first propose an overview of some basic measures of association, going from the simple co-occurrence frequency to the normative measure of contingency, r_φ. We then report an experiment collecting judgments of word-likeness as a function of the relationship between the phonemes composing the rimes (VC). The contingency between Vs and Cs, as assessed by r_φ, was the best predictor of children and adult judgments. Surprisingly, the forward transitional probability (P(C/V), which is the main measure considered by language researchers, was a poor predictor of performance, whereas the backward transitional probability (P(V/C)) made a sizeable contribution. We then analyze the ability of computational models to account for these results, successively considering a connectionist model based on the automatic computation of statistical regularities (SRN) [Cogn. Sci. 14 (1990) 179] and a model in which the sensitivity to statistical regularities emerges as a by-product of the attentional processing of the incoming information (Parser) [J. Memory Language 39 (1998) 246]. Somewhat ironically, Parser, which implements no specific mechanisms for statistical computations, proves to be a better predictor of performance than the SRN. The generality of these results, and their implications for the issue of automaticity, are discussed.

Keywords: Language; Phonology; Learning; Statistics; Correlations; Computational models

* Corresponding author.
E-mail addresses: pierre.perruchet@u-bourgogne.fr (P. Perruchet); ronald.peereman@u-bourgogne.fr (R. Peereman).
[1] http://www.u-bourgogne.fr/LEAD/

0911-6044/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0911-6044(03)00059-9

Since the early 1960s, the study of language is mainly related to the computational view of mind, in which symbolic information is processed through the application of formal rules. However, there is now growing evidence that people are also sensitive to the statistical regularities embedded into linguistic utterances. Because this sensitivity is necessarily acquired as a function of subjects' exposure to the language, such findings have led to a progressive decay of innatist views at the profit of theoretical perspectives centered on learning.

Many new questions arise as a consequence of this conceptual shift. The questions that will be addressed in this article are the following: what statistical regularities are learned by human subjects, and to what extent can their learning be thought of as automatic? In order to explore possible answers, let us consider an example taken from the domain of written language. In French, the phoneme /o/ can be transcribed with different graphemes, such as 'o', 'au', 'eau', 'ot', and so on. Which grapheme is used in a given case depends on various variables, such as the position of /o/ within words and the consonantic environment. For instance, in medial position, /o/ is more often spelled 'o' than 'au' between 'b' and 'r' but is more often spelled 'au' than 'o' between 'p' and 'v'. Likewise, in final position, /o/ is frequently transcribed 'eau' after 'r' or 't', but is never transcribed 'eau' after 'f'. Experimental data (Pacton, Fayol, & Perruchet, 2002) show that children become progressively sensitive to these graphotactic regularities. For instance, when asked to write non-words such as /bitavo/ and /bylefo/, children used the grapheme 'eau' more often in the former case than in the latter. This context effect was obtained as early as Grade 2, and increased further with grade level (6% in Grade 2, 27% in Grade 3 and 45.8% in Grade 4). Similar results were reported for other orthographic regularities in French (Pacton, Perruchet, Fayol, & Cleeremans, 2001) and in English (Cassar & Treiman, 1997).

Such learning can be labeled as automatic in several respects. It occurs without any directed teaching. This kind of regularity is not mentioned in textbooks, and is never pointed out to children by teachers or parents. Moreover, it is not a part of children's, or even adults' explicit knowledge. When people are questioned about the existence of phonotactic or graphotactic regularities (e.g. can /fo/ be spelled 'feau' at the end of words?), their only available strategy consists in mentally scanning a few relevant words (those endings by /fo/ in our example) in order to surmise a response. Irrespective of the success of this strategy, it appears manifest that people do not use it when asked to spell /bitavo/ and /bylefo/. In other words, performance is not mediated by the intentional exploitation of explicit knowledge. These properties allow an assertion that the sensitivity to statistical regularities is the product of what is now coined 'implicit learning' (for overviews, see: Berry & Dienes, 1993; Cleeremans, Destrebecqz, & Boyer, 1998). But this is the point at which consensus ends and a far-reaching conceptual bifurcation is opened.

The above depiction suggests that children are endowed with an ability to automatically compute statistics when exposed to a linguistic corpus. Children would perform various statistical analyses of the raw data without being consciously aware of the computations that occur in their mind. Arguably, the assumption that unconscious computations follow the very same algorithms that a statistician would follow when faced with the same data is somewhat implausible. However, the development of connectionist models suggests an alternative, and more attractive solution. The discovery of statistical regularity is performed by the progressive tuning of the connection weights between units within multilayer networks, a structure that has some (accordingly remote) similarity with the structure of a human brain. The simple recurrent networks (SRN), initially proposed by Elman (1990), have been the most widely applied to sequential learning. SRNs are typically trained to predict the next element of sequences presented one element at a time, and are therefore particularly appropriate to develop sensitivity to the sequential structure of the language. The introduction of SRNs and other connectionist networks provides some amount of biological plausibility to the idea that statistical computations are performed in the human brain (although several connectionist algorithms, including the SRN, do not correspond directly to any standard statistical method, see Redington and Chater (1998)). With regard to our concern about automaticity, this conception entails that sophisticated computations take place in our minds without any attentional and conscious counterpart.

To many, this interpretation is mandatory. However, there is an alternate way of thinking about the behavioral sensitivity to statistical regularity. In this alternative conception, sensitivity to statistical regularities is not the result of statistical computations on individual elements, but is the by-product of the local representations of chunks of individual elements. Indeed, there are 'two ways of learning associations' (Boucher & Dienes, 2003), and the question is open whether chunking is an emergent property of statistical analyses, or a primitive process the result of which amounts to simulate statistical computations. Although the notion of chunk is not necessarily linked to the concepts of attention and consciousness, we will consider hereafter that a chunk brings together the elements that are at the focus of one's attention at a given moment, and hence constitutes the momentary content of one's phenomenal consciousness (Perruchet & Vinter, 2002) .[2] How does this alternative conception work in the above example? Instead of considering that some computations of the relations between, say, 'v' and 'eau' occur in people's minds, it can be assumed that the attentional processing of different spellings of /vo/ naturally leads to increasing familiarity with some spellings at the expense of others, as a consequence of elementary mechanisms of associative memory: the memory traces of the more frequent segments are strengthened whereas the less frequent segments are forgotten. This, and a few other very simple mechanisms have been exploited in Parser, a computational model initially devised to discover words from a non-segmented speech flow (Perruchet & Vinter, 1998).

These two broad conceptions of learning obviously lead to quite different views about the automaticity of behavioral sensitivity to statistical regularities. However, before exploring these implications in more detail in the general discussion, we have to examine whether these conceptions are equally powered to account for the data. This requires an identification of the precise nature of the distributional information to which human subjects are sensitive, an issue that has been surprisingly neglected to date. After an overview of some basic statistical measures of association, we will examine which of them are predictive of human performance in a specific context, namely the relation between phonemes at an intra-syllabic level. Then we analyze the ability of models to account for

[2] In fact, the following experimental and simulation studies, with the exception of the final discussion about automaticity, do not depend on this postulate, and hence (hopefully) remains of interest for the readers who do not endorse this view.

these results, successively considering Elman's (1990) SRN, which involves the (automatic) computation of statistical regularities and Perruchet and Vinter's (1998) Parser, which relies on the (attentional) chunking of the incoming information.

1. Statistical measures of association

Although it is obviously possible to envision relationships between three and more events, here we will be restricted to the case of two successive events, e1 and e2. To explore the various measures of the relationships between e1 and e2, let us consider a 2×2 contingency matrix (Table 1) where *a* stands for the number of e1–e2 co-occurrences, *b* for the number of occurrences of e1 followed by an event different from e2, *c* for the number of occurrences of e2 preceded by an event different from e1, and *d* for the number of events comprising neither e1 nor e2.

A first index of relationship is given by *a*, which represents the frequency of e1–e2 pairs. The co-occurrence frequency is obviously relevant to our concerns. In the simplified presentation of the rational underlying Parser in the introductory section, we alluded only to the relative frequency of 'feau', 'fo', 'fau', and other spellings of /fo/, to account for children preferences. However, the pure co-occurrence frequency is quite limited as an indicator of relationships between two events. A more relevant measure is provided by the conditional or transitional probability (TP). The TP, *P*(e2/e1), is the probability that e1 is followed by e2, and can be computed as

$$\mathrm{TP} = \frac{a}{a+b} \quad (a) \tag{1}$$

Many authors (Peña, Bonatti, Nespor, & Mehler, 2002; Saffran, Newport, & Aslin, 1996) consider only the TP when they intend to quantify an association, for reasons that are somewhat unclear. Indeed, TP as such provides only part of the relevant information. In order to obtain a more reliable indicator of association, the TP (*P*(e2/e1)) must be compared to the probability of e2 when not preceded by e1. There is a predictive relation between e1 and e2 when the probability of e2 is larger in the presence than in the absence of e1. The resulting statistic is Delta *P* (ΔP), which stands as:

$$\Delta P = \frac{a}{a+b} - \frac{c}{c+d} \quad (b) \tag{2}$$

Shanks (1995), for instance, considers that ΔP is the normative measure of contingency, insofar as causal prediction is involved. Indeed, this measure indicates how well e2 can be

Table 1
A contingency matrix

		e2	
		+	−
e1	+	*a*	*b*
	−	*c*	*d*

predicted from e1. It is analog to a regression coefficient when data are dichotomic instead of continuous.

However, the strength of an association may also be related to the backward relationship between e1 and e2. Let us consider the TP and ΔP when e1 has to be inferred from e2. These coefficients will be denoted as TR$'$ and $\Delta P'$, respectively,

$$TP' = \frac{a}{a+c} \quad (c) \tag{3}$$

and

$$\Delta P' = \frac{a}{a+c} - \frac{b}{b+d} \quad (d) \tag{4}$$

Talking about 'backward' relations is potentially misguiding, because it suggests either (a) that the order of occurrence of e1 and e2 is reversed, or at least taken as irrelevant (which is obviously incorrect) or (b) that these statistics are aimed at predicting a past event, hence measuring information devoid of any adaptive function. Certainly, this explains why these measures are virtually ignored in the current literature. In fact, TP$'$ and $\Delta P'$ can be thought of as measuring the degree to which e1 is the only predictor of e2. Assuming the forward TP fixed, the interest of considering the relations between e1 and e2 might still vary as a function of whether there are predictors of e2 other than e1. If, for instance, there are better or more salient predictors, it may be adaptive to ignore e1 to focus on more relevant events, irrespective of whether e1 carries some predictive information when it is considered in isolation.

It makes sense to conceive that the tightness of the association between e1 and e2 depends on both forward and backward relationships. Accordingly, the standard measure of correlation, Pearson r, measures the two-way dependency between e1 and e2. With dichotomic data, Pearson r is commonly called "r phi" (r_φ), and expressed as:

$$r_\varphi = \frac{ad - bc}{\sqrt{(a+b) \times (c+d) \times (a+c) \times (b+d)}} \quad (e) \tag{5}$$

Interestingly, r_φ can also be expressed as the geometric mean of forward and backward ΔP (see Appendix A):

$$r_\varphi = \sqrt{\Delta P \times \Delta P'} = \sqrt{\left(\frac{a}{a+b} - \frac{c}{c+d}\right)\left(\frac{a}{a+c} - \frac{b}{b+d}\right)} \quad (f) \tag{6}$$

Eq. (6) makes more clear than Eq. (5) the increasing sophistication of the possible measures of association between two events, insofar as the Eqs. (1)–(4) are clearly embedded into it. Of course, other intermediary measures can be considered between the simple co-occurrence frequency and r_φ, and there are also alternative measures of association, based on other metrics (such as Mutual Information, which also assesses bi-directional relations). However, we limit ourselves below to considering the measures above, which are clearly hierarchized along a common metric. Which of these measures is human behavior the more sensitive to?

2. Which association measure is the best predictor of human behavior? The case of intra-syllabic phoneme relationships

Somewhat paradoxically in view of the upsurge of distributional approaches in recent years, the number of studies aimed at comparing the relevance of different measures of association in language is quite limited (for an example, see the comparison between frequency and TP in Aslin, Saffran and Newport (1998)). As a rule, authors select a given coefficient (e.g. the TP in Peña et al. (2002)), r_φ in Treiman, Kessler, Knewasser, Tincoff and Bowman (2000), but without explicit justification, and without considering alternative measures. This state of affairs is damaging because, without special care in the selection of the material, these measures exhibit substantial correlations. Thus it is unclear whether the effect attributed to the selected measure (e.g. TP) is not due in fact to another, correlated measure (e.g. r_φ).

In this section, we report experimental data aimed at exploring a range of measures in the specific case of intra-syllabic relations. More precisely, we focus on the relationship between vowel (V) and consonant (C) in the terminal syllable of words. Focusing on VC biphones is justified by the fact that the analysis of linguistic corpus suggests that there are strong probabilistic constraints on the VC combinations in the rimes, at least in English (Kessler & Treiman, 1997) and in French (Peereman & Content, 1997; Peereman, Dubois-Dunilac, Perruchet, & Content, 2003). Consistent with the structure of the data, behavioral data also suggest that rimes have a special role in the behavior of children and adults (Bryant, MacLean, Bradley, & Crossland, 1990; Goswami, 1986; Jared, 2002; Treiman, 1994; Treiman, Mullenix, Bijeljac-Babic, & Richmond-Welty, 1995; Wise, Olson, & Treiman, 1990). For instance, Treiman et al. (2000) report that 2 and 4° grade children judge nonsense CVC syllables as more wordlike when the rime includes highly contingent vowels and coda (as assessed by r_φ).

The Brulex lexical database (Content, Mousty, & Radeau, 1990) identifies 170 different VC rimes in the French language. When computed over the whole set of VC rimes, the mean value of the measures of association between V and C is surprisingly low. For instance, the mean value of r_φ is 0.015, that is, close to zero. Does this mean that contingency information would be irrelevant? Two reasons suggest not. First, when the values of the marginal means of a contingency matrix analog to that of Table 1 are not equilibrated (as is typically the case for language), the maximum value of r_φ never reaches one. Overall, on the 170 VC rimes, the average of the maximum value that r_φ can take given the marginal distributions is 0.520. But there is a second, more important reason. Given that the sample of V and C is finite, a positive contingency for a given VC string entails more or less mechanically that the contingency for the other pairs embedding either V or C will tend to zero or become negative. And indeed, 92 out of the 170 r_φ, that is to say more than half of them, are negative. But the point of interest is that contingency is positive for the most frequent items. For instance, the mean r_φ for the 10 most frequent VC rimes is 0.318. For TP and TP′, the corresponding mean values are 0.421 and 0.483, respectively. Thus the sensitivity to r_φ and other measures of association is adaptive, insofar as it should allow learning of the most frequent rimes.

Table 2 (upper panel) reports the intercorrelations between the statistics presented above. The higher correlations (around 0.90) are between TP and ΔP on the one hand,

Table 2
The correlations in the matrix of the upper panel were computed across the whole population of 170 VC rimes ending words in French

	Frequency	TP	TP′	ΔP	$\Delta P'$	r_φ
Freq.	1					
TP	0.593	1				
TP′	0.573	0.255	1			
ΔP	0.51	0.941	0.335	1		
$\Delta P'$	0.519	0.334	0.898	0.468	1	
r_φ	0.715	0.687	0.652	0.776	0.783	1
Children	0.504*	0.382	0.495*	0.534*	0.548*	0.600**
Adults	0.561**	0.355	0.619**	0.603**	0.679**	0.733**

The correlations with children and adult performances, shown in the lower panel, were computed across the sample of 23 VC used in the experiment (* = <0.05; ** = <0.005).

and TP′ and $\Delta P'$ on the other hand. By contrast, the lower correlations are between statistics considering the relations in the forward and backward orders. For instance, the correlation between TP (P(C/V)) and TP′ (P(V/C)) is only 0.255. Overall, the correlations between the main measures, although substantial, appear sufficiently far from one to justify our project, namely to search for the measure that is the best predictor of performance.

The following experiment is inspired from Treiman et al. (2000). Children and adults were presented with pairs of nonsense syllables differing by their rime. They were asked to indicate which of two CVC syllables sounded the most like French words. The rationale is to put in relation the degree to which each VC biphone sounds natural with its statistical characteristics, as assessed by the different associative measures considered above.

2.1. Method

2.1.1. Participants

Forty one children and 18 adults participated in the experiment. Children were divided into two age groups ($M = 5.7$ years : $n = 20$; $M = 8.9$ years : $n = 21$). Each age group corresponded to one school level, the youngest children coming from the last kindergarten grade and the oldest children coming from the third elementary grade. Adults were students majoring in psychology.

2.1.2. Material

Twenty four VC rimes were selected from the set of legal French rimes. They were chosen in order to fill in at best a 2×2 matrix, defined by the crossing of the frequency of the VC rimes, and the internal contingency assessed by the r_φ between the two phonemes. Thus six VC rimes were both frequent and contingent, six rimes were frequent but not contingent, six rimes were rare and contingent, and six rimes were rare and not contingent. Frequency and contingency measures were computed from the Brulex database (Content et al., 1990). It is worth stressing that, due to the correlations between the frequency and the contingency of intra-syllabic components in natural language (Pearson's $r = 0.715$

in French), it was not possible to obtain a perfectly orthogonal design. This lack of orthogonality is not detrimental for our objective, because correlational method rather than ANOVAS will be used in the analysis. For the purpose of the experiment, each of the VC rimes of one given cell of the 2×2 matrix was paired with one rime belonging to each of the three other cells. Pairing of the rimes was performed in order to contrast the rime either on frequency or contingency (or both). This operation resulted in 36 pairs of rimes. Then a consonant was added before each VC rime, with the proviso that the resulting CVC syllables did not correspond to French words, and that the CV unit resulting from the addition of the initial C was neither frequent nor contingent. The same consonant was added to the two VC rimes of a given pair. Appendix B presents the VC rimes and the final to-be-compared syllables. Two lists of 36 pairs of syllables were built, which differed by the order of succession of the pairs, and by the order of the syllables within each pair. The material was recorded by a female native speaker of French and digitalized (16 bits, 44.1 kHz), then transferred on a DAT recorder.

2.1.3. Procedure

Participants were tested individually in a quiet room. They were instructed that they had to listen to a list of pairs of syllables, and to decide which member of each pair was the most wordlike in French. Participants were randomly ascribed one of the two lists. Each pair of syllables was displayed twice in succession. One trial began by a brief auditory warning signal, followed 500 ms later by the first syllable of the pair. The interval between each syllable of a pair was 900 ms. The interval between the two repetitions of the same pair was 1500 ms. The next sound signal occurred immediately after the participants' response.

2.2. Results

A score of preference for each syllable was computed as being the number of participants selecting this syllable within a pair, then the scores were pooled over the three syllables involving the same VC rime. Pearson correlations were computed across the VC rimes between participants' preference and the different statistical measures of association. One VC rime (/ég/, which occurs only in the French word 'Touareg') out of the 24 was removed from the analysis, because its pattern of associative measures was quite atypical.

Because the results for kindergarten and third-graders did not differ in any interesting way, the two groups of children were pooled, so that data were analyzed for children on the one hand, and adults on the other hand. Results are reported at the bottom of Table 2. Participant's preferences were significantly correlated with most measures of association. For both children and adults, the strongest correlation was with r_φ. Somewhat surprisingly given the usual focus on this measure when assessing the statistical structure of linguistic material, the standard, forward TP was the only measure to not reach significance at the conventional threshold ($p = 0.072$ for children; $p = 0.096$ for adults). Stepwise regression analyses showed that r_φ significantly improved predictions made on the basis of TP (F to enter, Children: 6.86, $p = 0.016$; F to enter, Adults: 17.85, $p < 0.001$). In fact, correlations tended to be stronger when the dependency measures were computed in the reverse order than in the standard way. However, r_φ also improved prediction made on the basis of TP′, although the augmentation was only marginally significant for children (F to enter, Children: 3.49, $p = 0.076$; F to enter, Adults: 6.69, $p = 0.018$).

The correlations of participants' preference with the frequency of the VC rimes were sizeable and significant. However, the strong correlation observed between frequency and r_φ across the whole set of VC rimes in French prompts us to be cautious in assessing the importance of co-occurrence frequency. In fact, when frequency and r_φ were entered together as predictors of participants' preferences into a multiple regression analysis, frequency lost any predictive value ($F < 1$ in Children and Adults). Of course, this does not demonstrate that frequency as such did not play any role in preference judgments, but results do not allow us to reject this interpretation. On the other hand, when frequency was entered first, entering r_φ improved prediction in a way that was significant in Adults (F to enter $= 9.91$, $p = 0.005$) and marginally significant in Children (F to enter $= 3.23$, $p = 0.087$).

2.3. Discussion

This experiment shows that the judgment of children and adults as to the wordlikeness of CVC nonsense syllables is strongly dependent on the degree of association between the phonemes composing the rimes. This result confirms the results of Treiman et al. (2000) in English, and more generally, the growing number of studies showing that subjects are sensitive to the phonotactic or graphotactic regularities of their language. However, this experiment also reveals that the various statistical measures of association are not equally good predictors of human judgments. Surprisingly, the forward TP (P(C/V) for a VC rime), which is often the only statistic to be considered, was the worse predictor among our sample of measures. Moreover, statistics taking into account the degree to which V has competitors for the prediction of C appeared more closely associated with performance than the standard measure. This stands true for TP′, and to a lesser extent, for $\Delta P'$.

Finally, consistent with the two-way dependency revealed by the significance of both ΔP and $\Delta P'$, r_φ, which is the geometric mean of the two coefficients, appeared to be the best predictor of performance, and added some predictive value when the other measures (taken individually) were partialized out. Although the correlations of frequency with participants' preferences were substantial when considered in isolation, the data were compatible with an interpretation positing that these correlations ought to be attributed to the association between frequency and r_φ. Note that this result, which somewhat undermines the importance often ascribed to the simple frequency, can hardly be imputed to a bias in our sample of syllables, since care was taken to include both low- and high-frequency items in our experimental material.

3. Testing models

How well are the two models presented in Section 1 able to reproduce these results? One strategy would consist in training models with the same linguistic corpus to which children were exposed. Here we adopt an alternative strategy, in which the models are trained with a miniature database, specially devised to address in proper conditions a limited number of selected issues. Data collected in the Section 2 show that, with respect to intra-syllabic phonemic relations, (1) children and adults are only moderately sensitive to the raw co-occurrence frequency, (2) they are as much—and even more—sensitive to the backward relationship between C and V as to the standard, forward, prediction that V allows about C, and (3) as a consequence, performance is the most sensitive to the two-way dependency between phonemes as assessed by the r_{φ} coefficient. In order to explore whether models are able to account for these results, we built a database in which the contrast between the different measures of association, and notably between forward and backward relationships, was maximized. In addition, this alternative approach allowed the elimination of a number a features that are certainly relevant for the processing of the VC rime in natural language, but which are of no concern here. For instance, VC rimes are commonly embedded into a continuous speech stream, but some occur before a natural segmentation cue (e.g. when they end a phrase), a feature that may help the formation of a VC unit. Also, rime processing may depend on the nature of its constituents. For instance, Peereman et al. (2003) showed that the relations between the onset and the vowel depends on the sonority contrast between C and V, and it is possible that similar effects occur for the rimes. Although interesting on their own, these effects could interact in unknown ways with the effects of the statistical relationships between C and V, which are of interest here.

The data used for the simulations are displayed in Table 3. Twelve pairs of characters, each character standing for a phoneme and each pair standing for a VC rime, were randomly ordered to form a sequence, and this procedure was repeated 100 times to form a continuous stream of 2400 phonemes. One rime (AB) was maximally associated, as expressed by all of our measures. Indeed, A is always followed by B, and B is always preceded by A. Another rime (ED) is the opposite: E can be followed by other consonants than D, and D can be preceded by other vowels than E. The other rimes lie somewhere between these extremes, in a design that allows an independent consideration of forward

Table 3
Material used for the simulations

		C is always preceded by the same V	
		YES	NO
V is always followed by the same C	YES	AB	ID ŒD UD YD OD
	NO	EJ EK EL EM EN	ED

The measures of association for each pair are shown in Table 4.

Table 4
Columns labeled a, b, c, and d, report the frequencies needed to fill in the cells of the contingency matrix shown in Table 1, for the 12 VC rimes presented in Table 3, and for 3 selected CV pairs generated when the rimes are placed in succession

	a	*b*	*c*	*d*	TP	TP′	ΔP	$\Delta P'$	r_φ	SRN	Parser
AB	100	0	0	2299	1.000	1.000	1.000	1.000	1.000	0.980	810.6053
ID	100	0	500	1799	1.000	0.167	0.783	0.167	0.361	0.990	469.5043
ŒD	100	0	500	1799	1.000	0.167	0.783	0.167	0.361	0.990	508.0013
UD	100	0	500	1799	1.000	0.167	0.783	0.167	0.361	0.990	500.3768
YD	100	0	500	1799	1.000	0.167	0.783	0.167	0.361	0.990	482.4033
OD	100	0	500	1799	1.000	0.167	0.783	0.167	0.361	0.990	492.0773
EJ	100	500	0	1799	0.167	1.000	0.167	0.783	0.361	0.160	476.4286
EK	100	500	0	1799	0.167	1.000	0.167	0.783	0.361	0.150	452.1058
EL	100	500	0	1799	0.167	1.000	0.167	0.783	0.361	0.160	457.355
EM	100	500	0	1799	0.167	1.000	0.167	0.783	0.361	0.160	414.2585
EN	100	500	0	1799	0.167	1.000	0.167	0.783	0.361	0.160	426.9071
ED	100	500	500	1299	0.167	0.167	−0.111	−0.111	−0.111	0.150	151.5972
DE	312	287	288	1512	0.521	0.520	0.361	0.360	0.361	0.520	134.3081
DA	57	542	43	1757	0.095	0.570	0.071	0.334	0.154	0.090	0.000
BE	57	43	543	1756	0.570	0.095	0.334	0.071	0.154	0.570	0.000

For each item, the table also reports the value of the different associative measures, and the results of the simulations with SRN and Parser.

and backward relationships. To see how this works, the reader is invited to examine Table 4. This table shows the values for *a*, *b*, *c*, and *d*, which are needed to fill in the contingency matrix depicted in Table 1, and the different measures of associations between the pairs of phonemes composing the material. Table 4 includes three pairs straddling the 12 original VC rimes, namely: (1) DE, which turns out to be more than three times most frequent than the original VC rimes, but only moderately contingent, and (2) DA and BE, which are the more frequent (after DE) among the inter-unit pairs. Note that the specific symbols composing the cells of Table 3 have been selected arbitrarily, except that, to facilitate reading, the symbols standing for V are vowels and the symbols standing for C are consonants (of course, the label of the items was irrelevant for the simulations).

3.1. Simulations with SRN

The SRN, initially proposed by Elman (1990) (see also Cleeremans, Servan-Schreiber, & McClelland, 1989) is one of the most influential connectionist models in the implicit learning literature and, to a lesser degree in language research. SRNs are typically trained to predict the next element of sequences presented one element at a time to the network and are therefore particularly appropriate to develop a sensitivity to sequential structure. The network is presented, at each time step, with element t of a sequence, and with a copy of its own internal state (i.e. the vector of hidden units activations) at time step $t-1$. On the basis of these inputs, the network predicts element $t+1$ of the sequence. The network's prediction responses are compared with the actual successor of the sequence,

Table 5
Correlations between the predictions of SRN (first line) and Parser (second line) and the statistical properties of the pairs of phonemes, as described by the different measures of association

	Frequency	TP	TP′	ΔP	$\Delta P'$	r_φ
SRN	0.571	0.999	0.090	0.967	0.215	0.690
Parser	0.630	0.703	0.649	0.794	0.745	0.894
Children	0.504*	0.382	0.495*	0.534*	0.548*	0.600**
Adults	0.561**	0.355	0.619**	0.603**	0.679**	0.733**

Correlations were computed across the 60 pairs of phonemes generated when the material displayed in Table 3 is presented in succession. For the sake of commodity, the correlations obtained in the experiment and reported in Table 2 are also reprinted (lower panel).

and the resulting error signal is then used to modify its connection weights using the back-propagation algorithm.

To assess the SRN's ability to capture our data, we trained an SRN using Tlearn (Plunkett & Elman, 1997). There were 14 units in both input and output layers (one unit for each of the phonemes displayed in Table 3) and 20 hidden units. Unit activation ranged from 0 to 1, and the learning rate was set to 0.01. The whole database of 2400 phonemes described above was presented during each epoch. The network was trained for a total of 400 epochs. Network performance for the rime AB was assessed as the activation of the output unit corresponding to the phoneme B after the network had been presented with A, after averaging across all the occurrences of AB in the last epoch, and likewise for the other pairs.

The main results are shown in the penultimate column of Table 4. It appears that SRN's predictions match nearly perfectly the forward TP, P(e2/e1). This stands true for the selection of pairs shown in Table 4, but also for all the pairs generated in the sequence ($N = 60$), so that the correlation between the two measures reached 0.999, as reported in Table 5 (first line). Tables 4 and 5 also show that SRN's predictions are closely related to ΔP, a result that is hardly surprising, given that TP and ΔP are themselves related. By contrast, SRN's predictions were only weakly correlated with the TP′ and $\Delta P'$. Finally, the correlation of SRN's predictions was 0.571 with the co-occurrence frequency, and 0.690 with r_φ.

Overall, the correlations of SRN's prediction with both frequency and r_φ fit actual subjects' performance reasonably well. However, the network was exclusively sensitive to the forward relationship between V and C, whereas human participants were also, and still more, sensitive to the backward relationship between C and V. The mismatch between simulated and actual performance was not due to the choice of specific simulation parameters: although the full range of possible variations was not explored, the same pattern was obtained using different numbers of hidden units and a different learning rate. Furthermore, irrespective of human performance, it is worth stressing that the SRN is quite poor at discovering the original VC units composing the material. If one considers that the probability of formation of a given e1–e2 unit is proportional to the output activation of e2 when e1 is presented, it appears that the network created many CV units before the correct VC units. In fact, exploiting the network's results to segment the material leads to a formation of all the units ending with E (which straddle the initial units) before discovering those of the initial units that begin with E.

3.2. Simulations with Parser

Parser is based on the postulate that the sensitivity to the distributional structure of the material emerges as the natural by-product of the attentional processing of the incoming information. Only basic principles of associative learning and memory are involved. The primitives that are perceived within one attentional focus as a consequence of their spatial or temporal proximity are assumed to become the constituents of a new representational chunk. If the same primitives are no longer perceived together, the representation of the chunk rapidly vanishes, as a consequence of both natural decay and interference with the processing of similar material. However, if the same percept re-occurs, the representation of the chunk is progressively strengthened, and serves to guide further perception.

Originally, Parser was developed (Perruchet & Vinter, 1998) to deal with the situations proposed by Saffran et al. (1996), in which participants have to find the words forming an artificial language presented as an unbroken sequence of syllables. Its application to the present issue raised no special problem. The only changes brought to the version used in Perruchet and Vinter (1998) were the following: (1) the syllables, which served as primitives in the original version, were replaced by phonemes, (2) the rate of forgetting and interference were both set to 0.025, in order to deal with the new material,[3] and (3) although in the original version, perception was guided by the longest representational unit available in memory whenever several candidate units were possible, perception was shaped here by the shortest unit. Although this change does not prevent the formation of long units if long units turn out to be frequent enough, it favors the emergence of two-phoneme units, which are of primary concern here. The remaining parameters and the whole algorithm were left unchanged, so the reader may refer to Perruchet and Vinter (1998) for precisions about the characteristics of the model.

Parser received as input exactly the same material that was processed by the SRN, namely the corpus of 2400 phonemes, except that learning was stopped after five epochs instead of 400. In Parser, each item in memory is ascribed a weight, which increases when the item is perceived and otherwise decreases. The weights were taken to be the dependent variable (note that their absolute value is arbitrary, only their relative value is relevant). Table 4 (rightmost column) gives the weight of each biphone at the end of training, after averaging across 20 runs. Although the weights obtained by Parser correlated substantially with those provided by the SRN ($r = 0.694$, $N = 60$), there were also striking differences. The AB unit, which is the most contingent, had the highest weight. Following a decreasing order along a contingency scale, the following units were those in which e1 has several successors, or e2

[3] Because the rates of forgetting and interference depends on the aspects of the material that the model is not designed to detect (e.g. the degree of similarity between-items), these values need to be adjusted for each study. As in most computer simulations, the present values were those that provided the best fit with expected performance (after a cursory exploration of the parameters space). Although not fully satisfactory, this method provides at least an existence proof that expected performance are included in the solution space of the model. It can also serves as an heuristic tool of research, as pointed out later in the text. For a discussion, and a better method to deal with this issue, see Boucher and Dienes (2003).

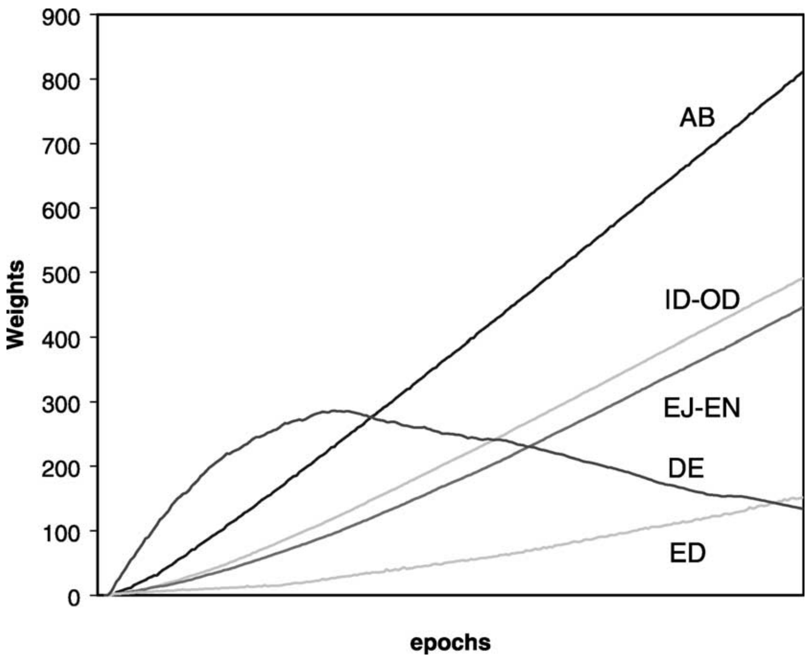


Fig. 1. Changes over training of the weights ascribed to the different pairs of phonemes by Parser. The pairs DA and BE, as well as all the other inter-rime pairs, are not represented, because they were absent from model's memory.

has several predecessors. The weights of the two categories of items were approximately equal. Finally, ED, which is the least contingent of the 12 initial units, had the lowest weight. As it appears in Table 5 (second line), which shows the correlations computed over all the units, this pattern translated into a strong correlation of the weights with r_φ, and correlations of approximately equal size with TP and ΔP according to their directionality.

It is noteworthy that Parser proved to be very good at discovering the initial units. Those units were discovered very early in training, and only those units were discovered with the exception of DE (i.e. the weight of the other pairs straddling the initial units was zero). In fact, the weight of DE was lower than those of the initial units. This result was unexpected. Why, for instance, was the weight of DE lower than the weight of ED, given that DE is more frequent than ED (312 vs. 100), that both forward and backward transitional probabilities are higher for DE than for ED (0.521 vs. 0.167 in both cases), and that DE is more contingent than ED ($r_\varphi = 0.361$ vs. -0.111)? To shed light on this question, we look at the dynamic of the learning process. Fig. 1 reports the evolution of the weights throughout the training process. It appears that DE was discovered early in practice, and kept the highest weight for a time. However, further training led to gradual vanishing (when training was pursued, DE disappeared from the model's memory), while the weight of the actual units went on to increase. We postpone to the general discussion the explanation of this astonishing performance.

3.3. Comparing the predictive power of SRN and Parser

The results of the above experiment are reprinted in Table 5 for comparison with the models' predictions. None of the models provided a close approximation. However, comparisons of the overall patterns of correlations at a qualitative level leads to the conclusion that Parser provides a better fit than the SRN. Indeed, Parser, like human participants, is sensitive to the two-way dependency between e1 and e2, and not only to the TP as usually computed (P(e2/e1)). As a consequence, both Parser's and actual participants' performances are essentially determined by the contingency, as measured by r_{φ}.

4. General discussion

We first described several measures of association clearly hierarchized on a dimension of complexity and sophistication, going from the simple co-occurrence frequency to the normative measure of contingency, r_{φ}. Then we explored which of these measures are the most predictive of human behavior, using the support of sensitivity to the VC terminal endings of words in French. Our experiment showed that children and adults felt that syllables sound like French words the higher the contingency between V and C. Surprisingly, the forward TP (P(C/V)), which is the main measure considered by language researchers, was a poor predictor of performance, whereas the backward TP (P(V/C)) made a sizeable contribution.

The ability of different models to account for these results was assessed through a comparison of an SRN, which is the connectionist network the most frequently used to simulate performance in sequential language tasks, with Parser, a computational model implementing the idea that the sensitivity to statistic regularity is the by-product of the natural attentional processing. Somewhat ironically, Parser, which implements no specific mechanisms for statistical computations, proved to be a better predictor of performance than the SRN, which was devised to learn statistical structure. In the following, we will discuss the generality of these results, dealing with empirical data and the models in turn, and then we will conclude on the implications of this study for the issue of automaticity.

4.1. Can be our empirical results generalized?

The fact that our data were collected on large samples of children and adults gives reasonably good assurance for their robustness in so far as intra-syllabic relationships are concerned. However, one may question whether the role of contingency, and the role of both forward and backward TP are fairly general over tasks and material. Our results appear to be partially at variance with those obtained in other areas of research. For instance, Shanks (1995) suggests that ΔP is the best predictor of performance in a variety of associative learning paradigms. This apparent discrepancy may support the distinction that Shanks draws between causal and structural relationships (Shanks, 1995, p. 2). Associative learning paradigms such as classical or operant conditioning could involve causal relationships, whereas the intra-syllabic relationships between phonemes, despite their temporal succession, could be structural in nature. It is also possible that the role of the backward TP has been underestimated, due to an exclusive focus on forward relationships, the use of which seeming to be, a priori, more rational. For instance, well-documented phenomena such as cue competition, blocking, and overshadowing, suggest that the degree to which a given conditioned stimulus is the only predictor of an outcome is quite relevant in the context of assessing the strength of a relation in the conditioning area.

Regarding more specifically the domain of language, the nearly exclusive focus on one, or at best two measures of association, does not allow any firm conclusion. For instance, the studies conducted by Aslin et al. are often taken as providing a compelling case for the role of forward TP. In fact, these studies take care of disentangling TP and relative frequency, but they do not envision the role of other statistical measures. Let us consider for example the material used in Saffran et al. (1996). The mean forward TP is 0.675 for intra-words pairs of syllables and 0.137 for between-words pairs, which means that this information may be used for the discovery of words, as Saffran et al. pointed out. However, the mean backward TP are, respectively, 0.655 and 0.146, which means that this measure is almost as predictive of words boundaries as forward transitional probabilities. The same is true for forward and backward ΔP and for r_{φ} (in each case, the coefficients for intra-words pairs of syllables were above 0.60, whereas the coefficients for between-words pairs were lower than 0.12).[4] Of course, the fact that a piece of information is available in the data does not imply that it is exploited by the subjects. Pending further studies, we are limited to speculation. It could be argued, for instance, that the relations between V and C would represent a special case on the ground that there are more V than C. In fact, at least in French, the number of V is nearly identical to the number of C (16 vs. 18, respectively). Other arguments, however, are more difficult to rule out. For instance, the succession of V and C in a diphone occurs in very close time succession. This may promote the formation of bi-directional associations, and hence, the role of r_{φ}, at the expense of the forward TP. By contrast, the associations between syllables or words, which are more remote in time, could be more sensitive to directionality. Further studies are needed to explore whether these factors are powerful enough to undermine the generality of our experimental results, or whether our main conclusions can be applied to other areas of language, such as word segmentation, and beyond, morphology, word classes, phrase structure, and lexical semantics (to cite the domains explored by Redington and Chater (1998), in their overview on the role of distributional information in language).

4.2. About connectionist models

Our study confirms the nearly exclusive sensitivity of an SRN to the forward transitional probabilities. This does not mean that the quasi-identity we obtained between SRN's prediction and the simple TP between V and C is general. This result is due to the fact that, in the material used for simulations, no information was provided by the items preceding any given V, because the VC rimes were displayed in randomized order. If there

[4] Transitional probabilities were computed in analytic ways, whereas ΔP and r_{φ} were approximated in Monte-Carlo simulations. In all cases, computations were performed after eliminating the two pairs of syllables that occur both in intra-words and between-words positions (*bupa* and *babu*).

were some grammatical constraints in the succession of VC rimes, the network could have exploited the information about second- or higher-order dependency available in the material, and its predictions would have outperformed the calculation of the simple, one-order TP. However, we believe it to be a robust conclusion that in the hierarchical set of association measures presented in the first part of this article, an SRN is linked to an intermediary level that makes it relatively insensitive to the raw co-occurrence frequency on the one hand, and to more sophisticated measures of associations on the other hand (by ‘relatively insensitive’, we mean insensitive to the specific information brought out by these measures, once their correlations with forward TP have been partialized out).

SRN’s relative insensitivity to the row frequency of co-occurrence is certainly a feature inherent to any connectionist models, because it is due to the high level of interference characteristic of neural networks. Although this characteristic is generally construed as beneficial, it can also be detrimental in some occasions. For instance, Boucher and Dienes (2003) show that a non-connectionist model, the Competitive Chunking model developed by Servan-Schreiber and Anderson (1990), proved to be a better predictor of performance than an SRN in an artificial grammar learning task, thanks to its relative immunity to interference processes.

By contrast, the exclusive sensitivity of an SRN to one-way dependency, and its resulting incapacity to make predictions matching a bi-directional measure of contingency such as r_φ, cannot be taken as inherent to connectionist networks. Indeed, this feature is derived from the SRN’s algorithm. It is not due to the objective of an SRN per se, namely to make predictions on the next event, but instead to the fact that the algorithm makes predictions independent from whether the predicted outcomes are shared by a number of different inputs or not. In order to examine the capability of other connectionist networks to reproduce our results, and more generally to detect genuine contingency relationships, we explored how an auto-associator performs when exposed to the same corpus as that used above (Table 3). Data were entered as a succession of pairs of phonemes. The input and output layers of the auto-associator comprised 28 units (one for each of the 14 different phonemes reported in Table 3, with a different unit for each phoneme according to whether it was in first or in second position within the pair). The hidden layer comprised 14 units. The other parameters were set as for the SRN above. The network indeed learned AB, which is the most contingent pair, better than most of the other pairs. However, there were exceptions: ED and DE were learned still more rapidly, although ED is the less contingent among the words, and DE is a between-word unit. Of course, these exploratory results cannot be taken as conclusive. We are quite confident that other networks can be designed to extract selectively the contingent items composing the material used in this article. The point that remains to elucidate is the extent to which these adapted networks will be specifically designed devices, or will be grounded on ubiquitous principles of cognitive functioning, as Parser is.

4.3. Why is Parser a so good learner?

Parser (Perruchet & Vinter, 1998) was initially designed to account for the segmentation into words of the artificial languages developed by Saffran et al. (1996), in which a few words are displayed in random succession. Because simple frequency brought out substantial information about words in those languages, one might suspect that Parser is restricted to the exploitation of such information. As a case in point, Hunt and Aslin (2001) wonder about the ability of Parser to discover words in a frequency-balanced task. Although not directly addressing the word segmentation issue, the present study shows that the model is able to discover the units composing a continuous sequence, even though the material was frequency-balanced, on the basis of more elaborate statistics than co-occurrence frequency measures. Moreover, Parser was able to correctly eliminate a between-rimes pair that the exclusive consideration of statistics should have lead to select.

The sensitivity of Parser to contingency relationships as indexed by r_φ, which is the normative measure of association, is all but mysterious. It lies in the fact that the model simulates the action of interference mechanisms. For instance, whenever A or B are presented in other contexts as the unit AB, they interfere with AB, resulting in a decrement of the weight of AB. Now, if AB is strongly contingent, this means that neither A nor B will be frequent in other contexts, and hence AB will receive no, or only a small amount of interference. Conversely, if AB is not (or negatively) contingent, this means that A and B are frequent events out of the AB unit, and therefore, interference will strongly reduce the weight of AB. Interestingly, modulating the parameter of interference makes Parser more sensitive to the contingency between events than to the relative frequency of co-occurrence (when the interference value is high) or more sensitive to the co-occurrence frequency than to the contingency (when the interference value is low). It is even possible to make the model exclusively sensitive to forward relationships, as an SRN (the interference should be limited to the A event) or to backward relationships (the interference should be limited to the B event).

Arguably, if the regulation of the interference parameter is purely ad hoc, without any possible external justification, the resulting good fit with people performance is of minor interest, and the model incurs a risk of unfalsifiability. The ideal situation is certainly when the level and the nature of interference can be set on a priori bases. Unfortunately, this situation may be the exception rather than the rule. Although the process of interference has been extensively studied in the literature on learning and memory, predicting when interference will and will not occur is not an easy task. However, the justification can also be provided a posteriori. If it turns out that the statistics to which humans are sensitive vary as a function of the tasks and the materials, the link that Parser suggests between the human sensitivity to various statistical measures of association and interference processes offers a potential avenue of investigation. Assuming for instance that further studies confirm that syllabic or sub-syllabic units are based on contingency whereas word segmentation is based on forward TP, it should be possible to examine whether interference processes operate differently in both cases through independent methods.

A somewhat more surprising ability of Parser was the elimination of a between-rimes unit (DE), despite the fact that this unit was the most frequent pair of all the material, and compared favorably to many of the rimes with regard to the other statistics? In fact, Parser is not sensitive to the distributional properties of the material in the same mechanical way as an SRN. As in Brent and Cartwright (1996) model, unit extraction depends on the overall consistency of the segmentation, because the units, which stands for attentional chunks, are discrete and disjunctive. Although DE is first considered as a relevant unit, its consolidation is precluded (Fig. 1), because it is not compatible with a convenient segmentation of the adjacent material. This property of Parser expands considerably its relevance for natural language, because it is the case that, in natural language, intra-word constituents are more closely associated than inter-word constituents, *but only on the average* (e.g. the phonemes corresponding to between-word clusters such as /so/, present in the expression 'piece of', are certainly more closely associated than those composing a number of rare words, as for example /sd/ in 'disdain').

To conclude, it should be misguided to consider that because Parser implements simplistic principles, it is underpowered whenever the information contained in the input becomes more sophisticated than raw frequency. On the contrary, as far as the situation investigated here is concerned, Parser turns out to exploit the relevant information in a much more clever way than the connectionist networks that we have explored as yet. This outcome confirms that Parser is not limited to deal with Saffran et al.'s (1996) artificial language for which it was initially designed, and extends further Parser's ability to learn from various situations, as revealed in recent studies (Peereman et al., 2003; Perruchet, Galland and Peereman, submitted; Perruchet, Vinter, Pacteau, & Gallego, 2002).

4.4. About the automaticity of statistical computations

As pointed out in the introductory section, the sensitivity to phonotactic or graphotactic regularities is unquestionably the end-product of implicit learning processes in several respects. Learning is incidental, and the regularities which people are sensitive to cannot be verbalized, except if they are the target of time-consuming and costful analyses of the contents of one's memory. These two criteria are those that are commonly retained to define implicit learning (e.g. Cleeremans et al., 1998). However, this does not resolve the question of knowing the very nature of the mechanisms involved in learning distributional information.

The most common conception consists in attributing the discovery of statistical patterns to mechanisms analogue to those that are implemented in connectionist networks. In other words, statistical regularities are discovered through the involvement of algorithms that have been specially designed to learn statistical regularities. For instance, an SRN is designed to learn the next element of a sequence, and in order to do that, its weights are modified through back-propagation as a function of the errors made in earlier predictions. The involved algorithm is ad hoc, in so far as there is no independent evidence that actual neural systems are modified through backpropagation (as acknowledged, e.g. by Elman et al. (1996)). More importantly for our concern, the mechanisms recruited for performing statistical analyses are fully automatic. They are unrelated to attention and to conscious experience.

We submit that the automatic computation of statistical measures is not a mandatory postulate. That is not to say that the computation of statistics is the endeavor of controlled or conscious processes. Our proposal is that there is no computation at all. Accordingly, Parser performs arithmetic operations. Although they are limited to a few additions and subtractions, it may be argued that, by the fact, Parser does not differ from other computational models such as connectionist networks. In fact, the point is not the degree of sophistication of the mathematical operations, but their objective. To draw an analogy, predicting the depth at which a hammer-stroke will drive a nail in a piece of wood may involve complex mathematical operations; However, nobody, we guess, would claim that the actual effect of the hammer is the result of a computational process. This is because the objective of computations is to simulate processes that are physical, not computational in nature. Likewise, the objective of the operations performed by Parser consists in mimicking with a digital computer the results of all-purposed biological mechanisms, the existence of which is independently asserted. Indeed, the only mechanisms involved in Parser are the strengthening of traces with repetitions, and their forgetting through decay and interference.[5] Arguing that strengthening and forgetting processes, which can be observed in any living organisms, involves computations, leads to deprive the word 'computation' from any informative content.

In one sense, our proposal reiterates the story surrounding the status of the rules, which evolved over recent decades. From the observation that human behavior is consistent with the general rules governing language and other materials, most investigators concluded that subjects extracted and applied those rules. Because these operations were not available to conscious thought, the concept of cognitive unconscious was created to account for rule extraction. Then it becomes increasingly obvious, in many domains such as reasoning, implicit learning, or language, that one must distinguish between a system that follows rules from one that simply conforms to rules (to borrow the distinction proposed by Smith, Langston, and Nisbett (1992)). A ball falling on the ground conforms to the law of gravity, but does not follow this law. Straightforward evidence has been gathered in the last years to support the idea that sensitivity to rule-governed material may be a by-product of the extraction of statistical regularities. Now, in the same way that sensitivity to rules has been commonly attributed to an automatic rule-extraction device, the sensitivity to statistical regularities is commonly attributed to the automatic computation of statistics. The concept of cognitive unconscious remains central, even though its main function has changed. Our proposal is that, to paraphrase Smith et al., one must distinguish between a system that follows statistical regularities from one that simply conforms to statistical regularities. Human sensitivity to statistical structure is not a proof of statistical computation, because it can be more simply the by-product of the attentional processing of the incoming information.

[5] Our reasoning does not exclude that these ubiquitous biological processes evolved through natural selection, partly because they enabled the living organisms to adjust their behavior to environmental contingencies, which is essential for survival.

Acknowledgements

We would like to thank Padraic Monoghan and Zoltan Dienes for their valuable comments on an earlier version of this manuscript, and Aurelien Perruchet for his contribution to Appendix A. This work has been supported by a grant from the French Ministère de la Recherche (Action Cognitique: Ecoles et Sciences Cognitives), by the Centre National de la Recherche Scientifique (CNRS, UMR 5022), and by the Universite de Bourgogne.

Appendix A

The usual presentations of r_φ in statistic textbooks demonstrate its equivalence with the standard product–moment correlation when data are dichotomic, and point out the relation of r_φ with the statistic χ^2 ($r_\varphi = \sqrt{\chi^2/N}$). However, we are aware of no reference mentioning its relation to ΔP. Here is a way to derive the usual formula of r_φ (Eq. (5), see text) from Eq. (6), which define r_φ as the geometric mean of ΔP and $\Delta P'$.

$$r_\varphi = \sqrt{\Delta P \times \Delta P'} = \sqrt{\left(\frac{a}{a+b} - \frac{c}{c+d}\right)\left(\frac{a}{a+c} - \frac{b}{b+d}\right)}$$

The formula can be re-written as:

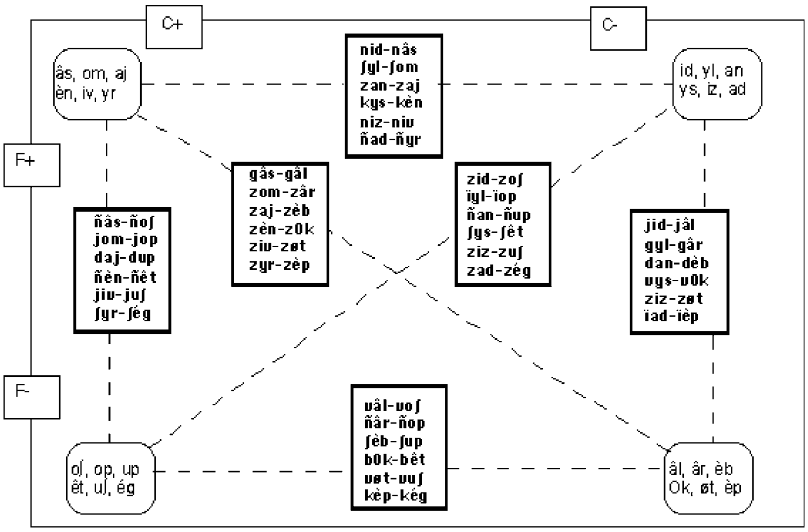
$$\begin{aligned}
&\sqrt{\left(\frac{a(c+d)-c(a+b)}{(a+b)(c+d)}\right)\left(\frac{a(b+d)-b(a+c)}{(a+c)(b+d)}\right)}\\
&=\sqrt{\left(\frac{ac+ad-ca-cb}{(a+b)(c+d)}\right)\left(\frac{ab+ad-ba-bc}{(a+c)(b+d)}\right)}\\
&=\left(\frac{\sqrt{ac+ad-ca-cb}}{\sqrt{(a+b)(c+d)}}\right)\left(\frac{\sqrt{ab+ad-ba-bc}}{\sqrt{(a+c)(b+d)}}\right)\\
&=\frac{\sqrt{a^2bc+a^2cd-a^2bc-c^2ab+a^2bd+a^2d^2-a^2bd-abcd-a^2bc-a^2cd+a^2bc+c^2ab-b^2ac-abcd+b^2ac+b^2c^2}}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}
\end{aligned}$$

The denominator is the denominator of the usual formula.
The numerator can be re-written as:

$$\sqrt{a^2d^2 - 2abcd + b^2c^2} = \sqrt{(ad-bc)^2} = ad - bc$$

Appendix B

VC Rimes and syllable pairs used in the Experiment (C + : highly contingent rimes; C − : not contingent rimes; F + : frequent rimes; F − : rare rimes)



References

Aslin, R. N., Saffran, J. R., & Newport, E.L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Berry, D. C., & Dienes, Z. (1993). *Implicit Learning: theoretical and empirical issues*. Hillsdale, NJ: Lawrence Erlbaum Associates, p. 197.

Boucher, L., & Dienes, Z. (2003). Two ways of learning associations. *Cognitive Science*, in press.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

Bryant, P. E., MacLean, M., Bradley, L. L., & Crossland, J. (1990). Rhyme, alliteration, phoneme detection, and learning to read. *Developmental Psychology*, *26*, 429–438.

Cassar, M., & Treiman, R. (1997). The beginnings of orthographic knowledge: children's knowledge of double letters in words. *Journal of Educational Psychology*, *89*, 631–644.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: news from the front. *Trends in Cognitive Sciences*, *2*, 406–416.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, *1*, 372–381.

Content, A., Mousty, P., & Radeau, M. (1990). Brulex: une base de données lexicales informatisée pour le français écrit et parlé. *Année Psychologique*, *90*, 551–566.

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

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plukett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Goswami, U. (1986). Children's use of analogy in learning to read: a developmental study. *Journal of Experimental Child Psychology*, *42*, 73–83.

Hunt, R. H., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General*, *130*, 658–680.

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language*, *46*, 723–750.

Kessler, B., & Treiman, R. (1997). Syllable structure and phoneme distribution. *Journal of Memory and Language*, *37*, 295–311.

Pacton, S., Fayol, M., & Perruchet, P. (2002). The acquisition of untaught orthographic regularities in French. In L. Vervhoeven, C. E. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy*. Amsterdam: John Benjamins.

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of the lab: the case of orthographic regularities. *Journal of Experimental Psychology: General*, *130*, 401–426.

Peereman, R., & Content, A. (1997). Orthographic and phonological neighborhoods in naming: not all neighbors are equally influential in orthographic space. *Journal of Memory and Language*, *37*, 382–421.

Peereman, R., Dubois-Dunilac, N., Perruchet, P., & Content, A. (2003). Distributional properties of language and sub-syllabic processing units. In P. Bonin (Ed.), *Mental Lexicon*. NY: NovaScience Publishers, in press.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*, 604–607.

Perruchet P., Galland, N., & Peereman, R. (submitted). Learning remote dependencies: no need for algebraic-like computations.

Perruchet, P., & Vinter, A. (1998). Parser: a model for word segmentation. *Journal of Memory and Language*, *39*, 246–263.

Perruchet, P., & Vinter, A. (2002). The self-organizing consciousness. *Behavioral and Brain Sciences*, *25*, 297–388.

Perruchet, P., Vinter, A., Pacteau, C., & Gallego, J. (2002). The formation of structurally relevant units in artificial grammar learning. *Quarterly Journal of Experimental Psychology, 55A*, 485–503.

Plunkett, K., & Elman, J. L. (1997). *Exercises in rethinking innateness: a handbook for connectionist simulations*. Cambridge, MA: MIT Press/Bradford Books.

Redington, M., & Chater, N. (1998). Connectionist and statistical approaches to language acquisition: a distributional perspective. *Language and Cognitive Processes*, *13*, 129–191.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

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

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge: Cambridge University Press.

Smith, E. E., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science*, *16*, 1–40.

Treiman, R. (1994). To what extent do orthographic units in print mirror phonological units in speech? *Journal of Psycholinguistic Research*, *23*, 1.

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. In M. B. Broe, & J. B. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon*. Cambridge: Cambridge University Press.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General*, *124*, 107–136.

Wise, B. W., Olson, R. K., & Treiman, R. (1990). Subsyllabic units in computerized reading instructions: onset-rime vs. postvowel segmentation. *Journal of Experimental Child Psychology*, *49*, 1–19.