Journal of Physiology - Paris 104 (2010) 19–26

Contents lists available at ScienceDirect

Journal of Physiology - Paris

journal homepage: www.elsevier.com/locate/jphysparis

An Adaptive Resonance Theory account of the implicit learning of orthographic word forms

H. Glotin [a,*], P. Warnier [a], F. Dandurand [b], S. Dufau [b], B. Lété [c], C. Touzet [d], J.C. Ziegler [b], J. Grainger [b]

[a] Laboratoire des Sciences de l'Information et des Systèmes (LSIS, UMR 6168), CNRS and Université du Sud-Toulon Var, BP 20132, 83957 La Garde, France
[b] Laboratoire de Psychologie Cognitive (LPC, UMR6146), CNRS and Aix-Marseille Université, 3 place V. Hugo, 13331 Marsellle, France
[c] Laboratoire d'Etude des Mécanismes Cognitifs (E.M.C), EA 3082, Université de Lyon / Université Lumière Lyon 2 / CNRS, 5 avenue Pierre Mendès France, 69676 Bron Cedex, France
[d] Laboratoire de Neurobiologie Intégrative et Adaptative (LNIA, UMR6149), CNRS and Aix-Marseille Université, 3 place V. Hugo, 13331 Marsellle, France

ARTICLE INFO

Keywords:
Neural networks
Adaptive Resonance Theory
Reading
Language
Modeling
IA
Cognitive Sciences
Dynamic Learning
Curriculum Learning
Visual Information to Semantics

ABSTRACT

An Adaptive Resonance Theory (ART) network was trained to identify unique orthographic word forms. Each word input to the model was represented as an unordered set of ordered letter pairs (open bigrams) that implement a flexible prelexical orthographic code. The network learned to map this prelexical orthographic code onto unique word representations (orthographic word forms). The network was trained on a realistic corpus of reading textbooks used in French primary schools. The amount of training was strictly identical to children's exposure to reading material from grade 1 to grade 5. Network performance was examined at each grade level. Adjustment of the learning and vigilance parameters of the network allowed us to reproduce the developmental growth of word identification performance seen in children. The network exhibited a word frequency effect and was found to be sensitive to the order of presentation of word inputs, particularly with low frequency words. These words were better learned with a randomized presentation order compared with the order of presentation in the school books. These results open up interesting perspectives for the application of ART networks in the study of the dynamics of learning to read.

1. Introduction

Skilled reading involves the ultra-fast and highly efficient mapping of low-level visual information about written words onto high-level semantic and syntactic representations. One central component among the multiple operations involved in the reading process is the parallel mapping of abstract letter identities onto whole-word orthographic representations (Rayner and Pollatsek, 1989; Grainger, 2008). It is this specific mapping process, and how it is learned, that is the focus of the present study.

Once the child has learned to map visual features onto abstract letter identities (see Grainger, Rey, and Dufau, 2008), one of the major challenges for the beginning reader is the processing of several letters in parallel. This processing of letters and letter combinations is thought to serve two purposes. Firstly, it enables the mapping of groups of letters (e.g., graphemes) onto phonology (e.g., phonemes), hence allowing the beginning reader to recover known phonological word forms and their associated meanings. This process is considered to be a key component in learning to read (Ziegler and Goswami, 2006). Secondly, given the clear information concerning word boundaries in printed text, it can be hypothesized that parallel processing of letter identities also provides a more direct, and therefore faster, route from visual information to semantics. This forms the basis of a generic dual-route approach to visual word recognition that has received much support in recent investigations of skilled reading (Coltheart, Rastle, Perry, Langdon, and Ziegler, 2001; Diependaele, Ziegler, and Grainger, in press; Perry, Ziegler, and Zorzi, 2007; Ziegler, Perry and Coltheart, 2003; Zorzi, Houghton, and Butterworth, 1998).

In the present study, we choose to focus on the unsupervised learning of orthographic word forms, which is at the heart of the direct route from visual information to semantics. The learning of orthographic word forms has largely been ignored up until now in computational investigations of the reading process. Many current computational models of visual word recognition are hardwired and therefore choose to ignore issues of learning in order to focus on issues of performance. Furthermore, most models that have chosen to deal with issues of learning have opted for the back-propagation algorithm (Harm and Seidenberg, 1999, 2004; Plaut et al., 1996; Seidenberg and McClelland, 1989). However, back-propagation is developmentally implausible because a significant part of the learning-to-read process proceeds without supervision (Share, 1995). Moreover, many studies observed that

---

* Corresponding author. Address: Université du Sud-Toulon Var, Avenue de l'université, B.P. 20132, 83957 La garde Cedex, France. Tel.: +33 04 94 14 28 24; fax: +33 04 94 14 28 97.
E-mail address: glotin@univ-tln.fr (H. Glotin).

0928-4257/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jphysparis.2009.11.003

relatively few exposures to a word are sufficient for the acquisition of word-specific orthographic information (for review, see Share, 1995). However, the above cited back-propagation models need hundreds of exposures during training in order to obtain satisfactory learning performance (e.g., Hutzler et al., 2004). Finally, human learning is sequential, and back-propagation models typically encounter serious problems when trained with realistic sequential input (see McCloskey and Cohen, 1989). This problem is referred to as *catastrophic interference*, which reflects the fact that training on a new set of items may drastically disrupt performance on previously learned items.

Given the above described shortcomings of previous models, the goal of the present study was to test a developmentally plausible model of orthographic word learning. The model should have the following features: (1) it should be able to learn in an unsupervised fashion; (2) it should be able to learn rapidly with only a few exposures to given word, and (3) it should be able to learn sequentially using realistic input (real text book materials). Because of these constraints, we turned to the Adaptive Resonance Theory (ART) as a potential candidate which might offer viable solutions to some of these problems.

1.1. The Adaptive Resonance Theory (ART)

In the present research, Adaptive Resonance Theory (ART) is applied as a method for the unsupervised learning of orthographic word forms. ART provides a general theoretical framework for learning and plasticity that spans the neural and cognitive levels. ART has given birth to several algorithms ART1, ART2, ART3, ARTMAP and many derivatives such as FuzzyART, SMART. It has been successfully applied as a model of low-level perceptual phenomena, as well as higher-level cognitive processes such as attention, memory, and recognition (Carpenter and Grossberg, 1986a,b), and speech perception (Grossberg et al., 1997).

The ART2 algorithm is applied in the present work because of its capacity to deal with analogue input code. Carpenter and Grossberg (1987) have demonstrated how the ART2 algorithm can solve the stability/plasticity dilemma, where the storage of new information does not systematically destroy previously learned information. ART networks have several attractive features for modeling cognitive-level phenomena. First, ART is sensitive to the order in which data are presented, unlike many other neural network models. Second, ART networks begin with minimal size, and expand as needed to learn a given task. Third, with ART, the stability–plasticity tradeoff can be explicitly controlled or manipulated. Capturing the inherent stability and plasticity of learning in biological systems, a challenge for many computational models, is ART's foremost quality. Furthermore, ART is a member of a class of unsupervised learning algorithms, so there is no external signal or information to learn from. Consequently, this class of models can only learn the statistical regularities of the data themselves. Even though it is simple, unsupervised learning is powerful and plausible, and biological systems are known to learn from mere exposure to stimuli.

1.2. The present study

The ART network was trained on a part of the Manulex corpus (Lété et al., 2004) which contains 1.9 million words taken from 54 readers used in French primary schools between the first and fifth grades (age 6 until 11). To mimic the learning of orthographic word forms through grade L1 to L5, one schoolbook was chosen by chance at each grade ("Gafi" and "Arthur"). In this study we keep only words of four and five letters. Thus there were 4872 words (858 different categories) in Grade L1 schoolbook, 6266 (875) in L2, 8164 (1076) in L3, in 7158 (1077) in L4, and 8243 (1282) in L5. The model was then trained, word by word, either in the correct order (in the main simulations), in random order and in two deterministic sequences in which words are presented in a grouped fashion, either in ascending or in descending order of frequency (in further tests of the model). Each word input was initially coded as an unordered set of ordered letter pairs (open bigrams), mimicking the type of approximate flexible coding revealed by empirical investigations of orthographic processing (Grainger, 2008). An ART2 algorithm was then used to adjust connection strengths between the input code and a higher-level code that learns to represent orthographic word forms (i.e., sequences of letters that appear between spaces).

There are several important novel aspects to the present work. First, we focus on unsupervised learning of orthographic word forms. Second, we use a biologically inspired learning algorithm (ART). Third, we use a realistic training corpus. Fourth, the present study provides an investigation of the importance of order of training on network performance.

2. Methods

2.1. The orthographic input code

A key question in current research on visual word recognition concerns the precise nature of the orthographic input code (for review, see Grainger, 2008). Although there is a general consensus that this code represents the identities of the word's component letters, there is less agreement on how information about letter position is coded (Goswami and Ziegler, 2006). Empirical research applying a variety of paradigms has provided evidence for the existence of an approximate, flexible, prelexical orthographic code for printed words, designed to optimize the mapping of visual information onto semantics via orthographic word forms (e.g., Grainger et al., 2006; Schoonbaert and Grainger, 2004; Perea and Lupker, 2004). There are various ways to implement this type of flexible prelexical orthographic code (e.g., Davis, 1999; Gomez, Perea and Ratcliff, 2008). In the present work, we opt for one particular solution, preferred by several investigators (Dehaene et al., 2005; Grainger and van Heuven, 2003; Whitney, 2001). This specific solution codes for combinations of letters in the correct order, without necessarily being adjacent – so called "open bigrams".

The input vector of our model represents every possible open bigram in French. This vector contains 1681 elements going from AA to ZZ, (41 orthographic signs (i.e., letters) in French including letters with accents). Elements along the vector code, in binary form, for presence (value = 1) or absence (value = 0) of the corresponding open bigram. All French words can be coded onto such 1681-element vectors of open bigrams.

For instance, open bigrams of Table are coded as follows:

AA	...	AB	...	AE	...	AL	...	BE	...	BL	...	LE	...	TA	...	TB	...	TL	...	ZZ
0		1		1		1		1		1		1		1		1		1		0

After having chosen open bigram coding, we further included plausible visual constraints on our coding. Here we apply the empirical results of a study investigating variations in letter visibility in five and seven letter words (Stevens and Grainger, 2003). These authors found that recognition probability of a letter depends on its position in a word. The values obtained for 5- and 7-letter strings were extrapolated in order to obtain values for all word lengths processed in our model. These coefficients were used to modulate the binary values of the input vector, as a function of the position-in-word of each individual letter of an input bigram. The activation value of a bigram is equal to the average recognition probability of its component letters.

In sum, an input for a word is a 1681-element vector of recognition probabilities of corresponding bigrams which both take into account their presence or absence, and also how their recognition probability is modulated by visual processing. One of the main properties of these input vectors are their low density; for example, a 10-letter word has only 27 non-zero elements. As we will see, this characteristic of inputs influences the choice of ART2 parameters.

2.2. ART implementation

The ART2 architecture consists of a superposition of two layers of representations, F1 the bigram level (i.e., a prelexical orthographic code), and F2 the lexical level (i.e., orthographic word forms), as represented in Fig. 1.

2.3. The learning (L) and vigilance (V) ART parameters

When ART is presented with a new data pattern, it can do one of two things: recognize it as a member of a category it already knows, or learn it as a new category if it is sufficiently different from anything seen so far. Each category is associated with a neuron. Thus in ART, recognizing a new category is synonymous to adding a new neuron. A parameter called vigilance (*V*) controls how different an incoming pattern must be in order to be considered as belonging to a new, unseen category. If vigilance is low, different input patterns tend to be grouped together using only a few (maybe even only one) very broad categories. At the other end of the spectrum, if vigilance is very high, incoming data patterns are always considered as coming from a different category, and thus all individual patterns are memorized, resulting in poor generalization. Intermediate values allow similar items to be classified as belonging together, while maintaining the ability to create new categories for items that are sufficiently different.

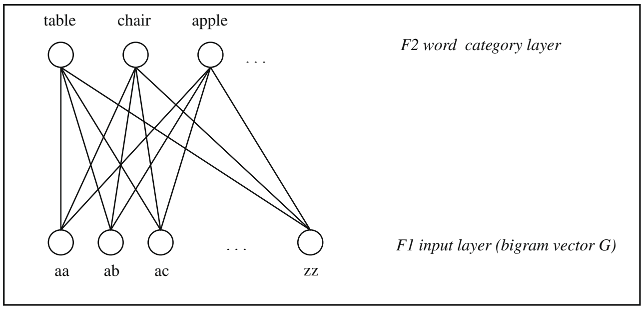

Fig. 1. Representation of the information stored in F1 and F2. Each unit in F1 is associated, by bottom-up and top-down connections with one open bigram. French has 41 distinct letters when including letters with accents, giving rise to 1681 (41 × 41) open bigrams. A word such as TABLE activates nine units in F1. In F2, each unit is associated with one or many word(s).

When a data pattern is recognized as belonging to a certain category, the weights of the associated category are modified by a certain amount in the direction of the newly presented data. These weights can be conceptualized as representing a *prototype* of the category. The amount by which weights associated with a given category are modified is controlled by a second ART parameter called learning (*L*). This is akin to the concept of learning rate in other models. A high value of the learning parameter can result in possibly fast learning, but might also be potentially unstable and non-converging. Lower values of the learning parameter improve convergence and stability, at the cost of a slower learning process.

2.4. The ART algorithm

The ART learning algorithm, with its topology represented in Fig. 2, is summarized below (more details can be found in Grossberg, Boardman and Cohen (1997)). The main connections are the bottom-up connections b_{ij}, from cells G_i to cells C_j, and the respective top-down connections t_{ji}.

Here is a summary of the ART algorithm:

Step 1. Initialization: all connections b_{ij} and t_{ij}, and F2 activations and d are set to zero.

Step 2. Get the current data sample, i.e., the bigram representation of the current word is the input vector pattern $\mathbf{G}$ in F1, with norm($\mathbf{G}$) = 1.

Step 3. If F2 is empty then create a category (step 5), otherwise activate the categories for $\mathbf{G}$ in F2 through bottom-up connections by: $y_j = b_{ij}\cdot G_i$, for all *j* in F2.

Step 4. Winner-takes-all competition: the candidate unit (called *J*), in order to learn the input pattern $\mathbf{G}$, is the one that maximizes y_j. Its activation is $d = \max(y_J)$. If $d > V$ (the vigilance parameter), go to step 5, otherwise go to step 6.

Step 5. Update weights for the candidate unit y_J: $t_{ji} = L\cdot G_i + (1 - L)\cdot t_{ji,}$ and $b_{ij} = L\cdot G_i + (1 - L)\cdot b_{ji,}$ for all *i*, where *L* is the learning rate.

Step 6. Repeat from step 2 for all input patterns.

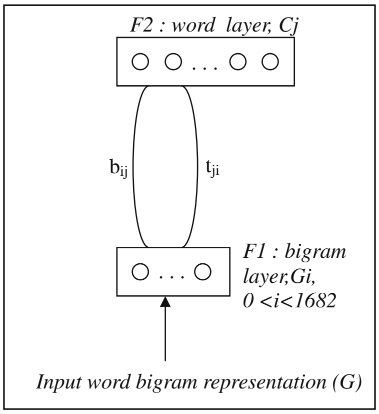

Fig. 2. The ART architecture with open bigrams (F1) and words (F2). Activities in F1 are the result of activities G_i of the word inputs, according to a winner-take-all competition and update of bottom-up and top-down connections. The number of recruited categories C_j in F2 varies from one to the number of presented inputs.

We process this algorithm only once, from the first to the last word instance of the sequence. Each word instance has an index position in the sequence, from 1 to the length T of the sequence. Thus for one grade level, we present T inputs to the network. After training, the learned categories are memorized with their activities in the F2 layer. The number of learned categories (C) can vary between 1 and T. A perfect learning is achieved if C equals the number of different words in the sequence. After training, we consider that a word is learned (or correctly activated), if its last instance in the sequence has not been recruited by another word. Thus, we can compute the percent of Correct Word Identification (CWI) for different sequences.

3. Results

The ART network model was trained using the same school books that children are exposed to in French primary schools. The mean percentage correct response for the model was given as the percentage of presented words that were correctly activated for their last presentation at each grade. We then analyzed the average percentage correct scores at each grade level for different V and L parameters.

3.1. Comparison with empirical children behavioral data

The trained network was evaluated against the data of an empirical study investigating the development of word identification accuracy from grade 1 to grade 5 in French primary school children. These children learned to read using the same textbooks that were used to train the model. The results of the behavioral study are taken from Dufau et al. (in press) and reported in Table 1. The study manipulated word frequency and word similarity. Word similarity is expressed in terms of orthographic neighborhood (words that differ from another by one letter).

Table 1
Behavioral results. Mean correct response percentages on words from grade 1 (L1 or "CP") through grade 5 (L5 or "CM2").

Grade level	HF–HN	HF–LN	LF–HN	LF–LN
L1 (CP)	84 (13)	79 (15)	54 (18)	46 (17)
L2 (CE1)	91 (7)	89 (13)	70 (16)	62 (17)
L3 (CE2)	96 (7)	94 (9)	78 (11)	74 (14)
L4 (CM1)	94 (7)	94 (7)	81 (10)	82 (12)
L5 (CM2)	99 (4)	98 (5)	84 (12)	85 (12)

Standard deviations in parentheses; HF, high frequency; LF, low frequency; HN, high neighborhood; LN, low neighborhood.

These behavioral scores are also shown in Fig. 3, along with the model's performance for a given combination of V and L parameters that allowed the model to reproduce the empirical learning curve. Given these satisfactory fits, we felt no need to overfit V and L parameters using LMS criteria. We get similar ART results in the ranges $0.85 < V < 0.92$ and $0.35 < L < 0.45$.

3.2. Testing ART's sensitivity to word order

In this section, we compare network performance following training with words presented in the order they appear in textbooks (this initial sequence is called the "SEQ" sequence in the following sections), versus training on the same word set but presented in random order ("RAND" sequence), or sorted sequences ("SORT" as defined in Section 3.4).

The random sequences were simply built as follows. We randomized, at each grade level, the order of words extracted from the corpus. Five random sequences were created for each level. The results are given in Fig. 4 for different L and V parameter settings. Comparing the SEQ and RAND word recognition performance over 12 networks for all the words, we see that RAND sequences significantly outperform SEQ and SORT sequences.

3.3. High frequency word learning is less sensitive to word order

A test was performed separately for different word frequency intervals. Ten intervals were computed, from word frequencies of −4 to −2 (in log10 scale). Fig. 5 shows the distribution of word frequency values and the corresponding Correct Word Identification rates of the model on SEQ and RAND sequences, averaging across all grade levels. This analysis demonstrates that word order has the biggest impact on network performance with low frequency words.

3.4. Learning on sorted word instances

The previous section has demonstrated an effect of word order on network learning by comparing randomized sequences with the

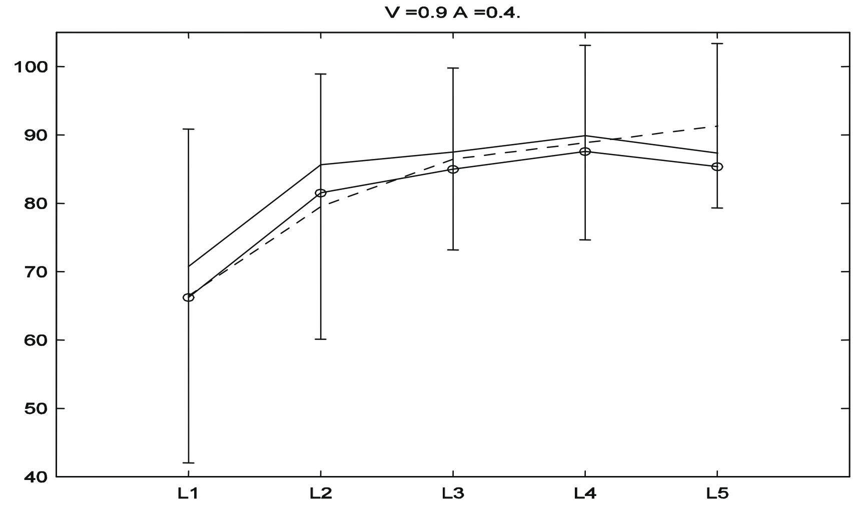

Fig. 3. ART simulation (at $L = 0.4$ and $V = 0.9$) results as percent of Correct Word Identification (CWI), over the five grade levels L1 to L5 (French primary education), resulting in performance matched to that observed in children that are also represented here in dashed '- -' with their std. ART trained on initial sequence 'SEQ' is represented by 'o-', and ART on RAND sequences by '__'.

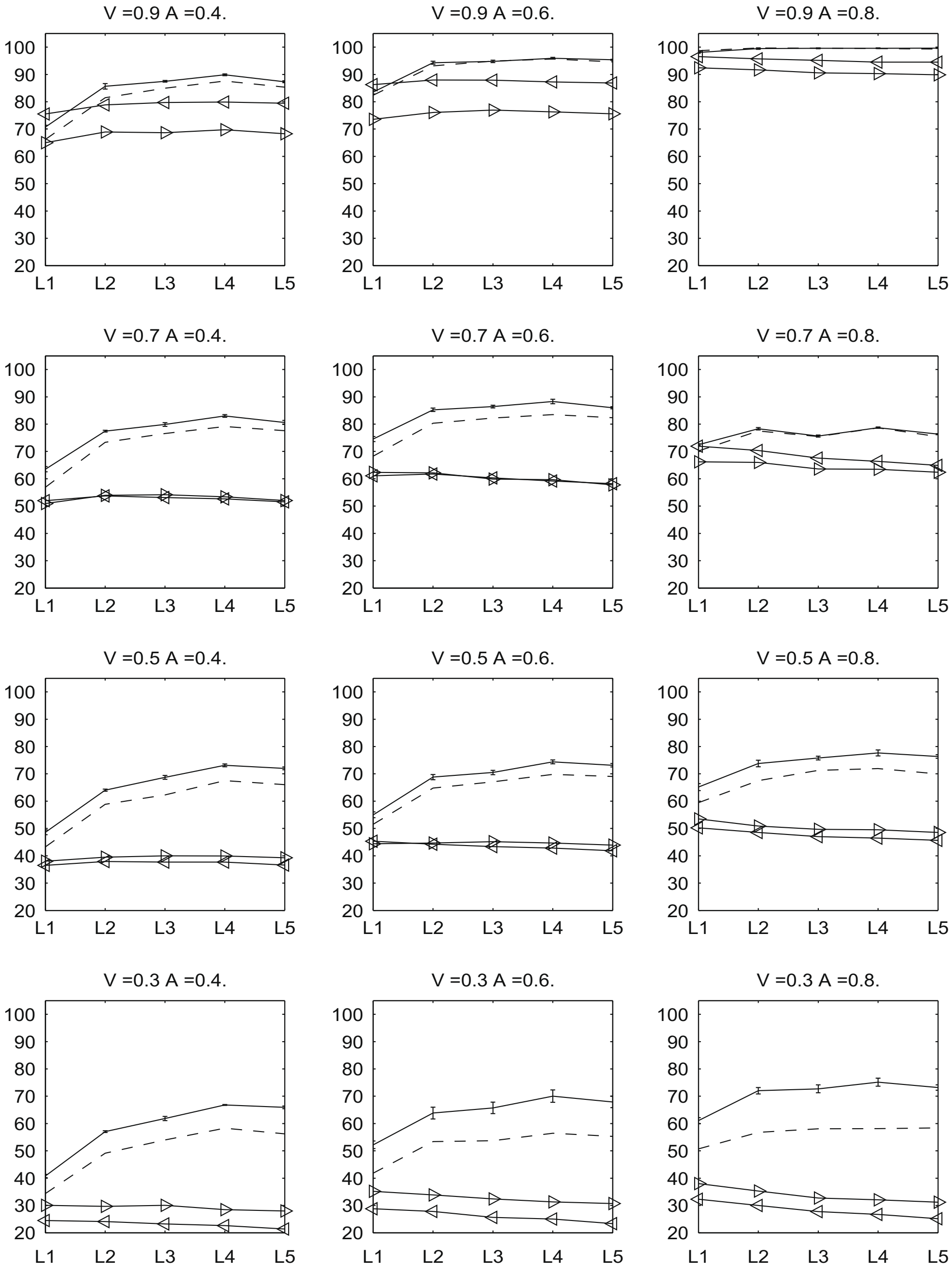

Fig. 4. Percent of Correct Word Identification (CWI) of ART model, from level L1 to L5, and for different *L* and *V* parameters, trained on the original sequence SEQ ('- -'), or on the five random sequences (average values '__' with std.), or on sorted SORT ('- > -') versus SORT REVERSE ('- < -') sequences. The upper left subplot is detailed in Fig. 3.

sequences found in school books. As a further investigation of the effects of word order, we trained the model with even more orderly sequences than found in the school book corpus. We chose to study the two extreme sequences. The first, that we call the sequence «SORT», involves presenting each word instance in only one shot.[1] The repeated word sequences are ordered from the shortest to the longest (i.e., from the lowest to the highest frequency), so we have: SORT = [$W_{1,1}$ $W_{2,1}$ $W_{3,1}$ $W_{3,2}$ $W_{3,3}$ $W_{4,1}$ $W_{4,2}$ $W_{4,3}$ $W_{4,4}$ $W_{5,1}$ $W_{5,2}$ $W_{5,3}$ $W_{5,4}$ ...], where $W_{p,q}$ is the *q*th instance (or sample) of the word W_p.

The second sequence is called «SORT-REVERSE». It is the same as the SORT sequence, but in reverse order, with high frequency words appearing before low-frequency words, so we have: SORT_REVERSE = [... $W_{5,4}$ $W_{5,3}$ $W_{5,2}$ $W_{5,1}$,$W_{4,4}$ $W_{4,3}$ $W_{4,2}$ $W_{4,1}$ $W_{3,3}$ $W_{3,2}$ $W_{3,1}$ $W_{2,1}$ $W_{1,1}$].

We see in Fig. 4 that the sequences SORT and SORT_REVERSE have nearly the same scores, with SORT better than SORT_REVERSE for low *V* (and vice versa). Except for high *V*, the network performs worse with all of these higher order sequences compared with the RAND and SEQ orders.

4. Discussion

In the present study we have shown that Adaptive Resonance Theory (ART) provides a viable framework for modeling the implicit learning of orthographic word forms. Our implementation of ART demonstrated rapid learning of orthographic word forms from

---

[1] We call a «shot» a continuous repeated presentation of all the instances of a word.

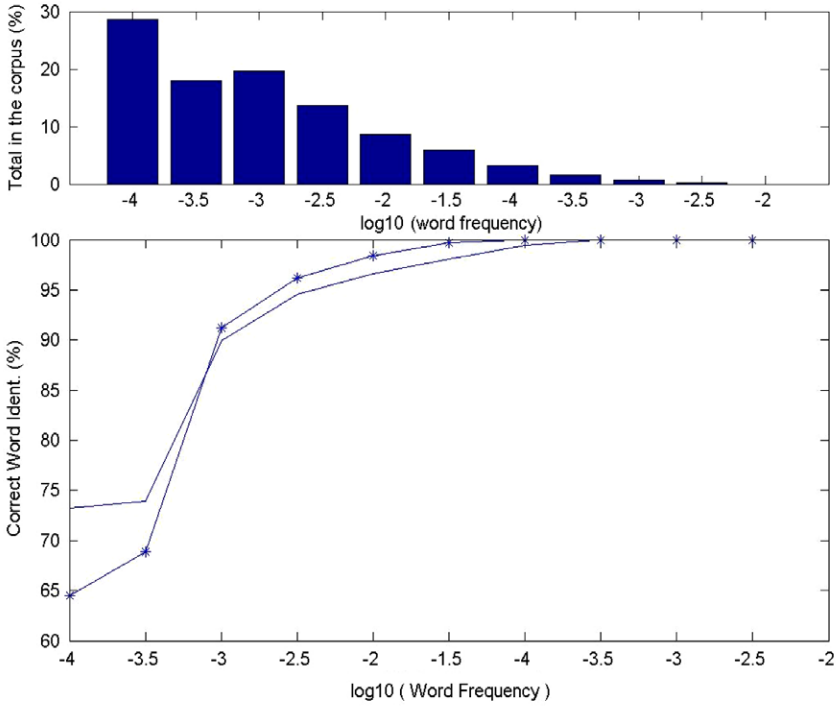

Fig. 5. Top global word frequency distribution across all grade levels and respective percent of Correct Word Identification (CWI) of ART model ($V = 0.9$, $L = 0.4$), on the original sequence SEQ ('-∗-'), or one random RAND sequences ('__'). Most of the word categories are infrequent. The CWI is higher for RAND than for SEQ for these infrequent words. Thus ART trained on RAND has in average a higher CWI than when trained on SEQ.

an approximate, flexible, prelexical orthographic input code (open bigrams – Grainger and van Heuven, 2003). Most important, the results of training the network on a realistic training regime (i.e., reading textbooks used in primary education) revealed a developmental pattern that mimicked that seen with children (Dufau et al., in press). Not surprisingly, the model showed two other key properties – a sensitivity to word frequency and a sensitivity to order.

The simulations show that ART reproduces the classic word frequency effect – the fact that words that occur more frequently in a given language are processed more rapidly and more accurately than words that occur less frequently (e.g., Balota and Chumbley, 1984; Rubenstein, Garfield and Millikan, 1970). More interesting, perhaps, is that in our ART network, speed and success of learning not only depended on frequency of presentation to the network, but also on the order of presentation.

Learning low frequency words was more successful when these words were presented to the network in a randomized presentation compared with the order they appear in the school textbooks. In fact, simulation results suggest ART performs best when words are regularly repeated instead of being presented in a grouped fashion. This effect is reminiscent of catastrophic interference where unseen items tend to be forgotten. Uniform random distributions tend to make words repeat regularly, and thus less likely to be forgotten. It is important to note however, that ART still retains some knowledge of words even in the extreme conditions (SORT and REVERSE SORT), and thus is still somewhat resistant to catastrophic interference.

4.1. Word dispersion analyses

In order to quantify the distribution of word instances, we analyzed, for each word, the entropy of the positions of its occurrences. A word position is its index in the sequence: an integer between 1 and the length of the sequence. The entropy is computed over 32 bins, thus $H \leqslant 5 = \log_2(32)$ bits. The results are given for each bit interval in Fig. 6, for sequences SEQ and RAND, in grade level L1. We see that $H \leqslant 1$ for 55% of the words of the SEQ sequence, compared to 43% for the RAND condition, with identical word frequencies. Respectively, RAND sequences have more words than SEQ with $H > 1$.

Similar results are obtained on other levels. They reveal that some word instances tend to be grouped together in the school

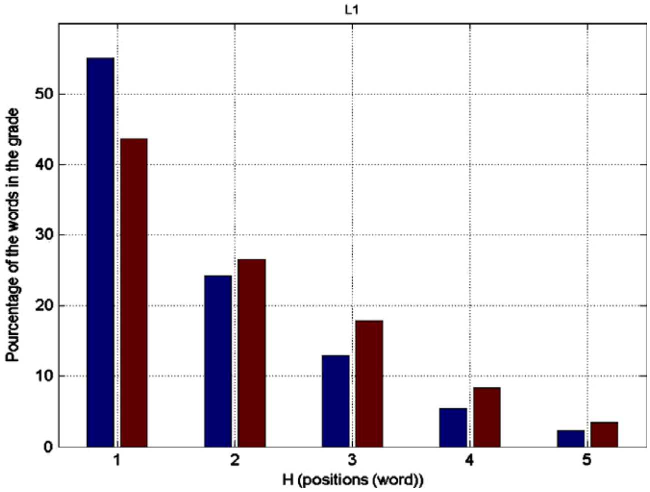

Fig. 6. Histograms of the entropy H of the positions of each instance of each word in the sequence SEQ (left blue bars), and in the five RAND sequences (right red bars), on grade L1. This entropy is computed considering the positions of all the instances of a given word over 32 bins, thus $H \leqslant 5$ (see details in Section 4.1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

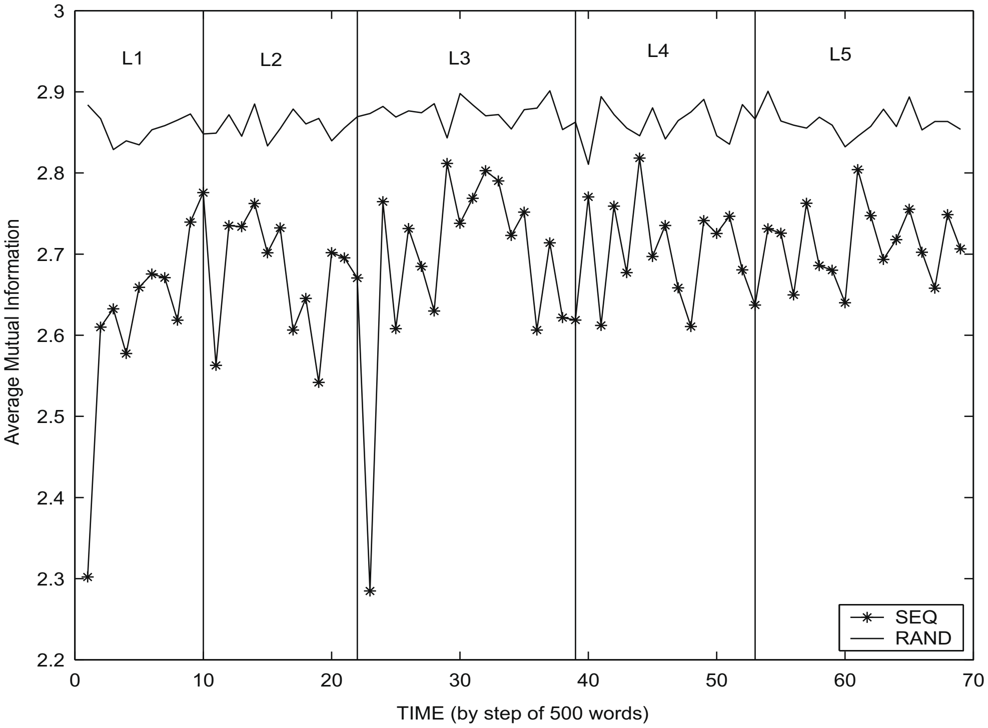

Fig. 7. Average Mutual Information AMI (S, time) for the sequence SEQ ('-*') and for one of the RAND sequence ('__') measured in bits. Vertical lines represent the end of each grade level. We see AMI(SEQ) < AMI(RAND) consistently and AMI(SEQ) is much more variable than AMI(RAND).

book corpus (SEQ), significantly more so than would occur by chance, such that repetition lag can be very large in certain conditions.

4.2. Sequence structure analysis by local mutual information

In this section, we investigate sentence structure using local mutual information. For each word W, the input vector G is the recognition probabilities of corresponding bigrams. In a window of n consecutive samples $[W_1, \ldots, W_n]$ of a sequence S, we can approximate local mutual information at position p as follows:

$$\mathrm{MI}_{[Sp,Sp+n]}(W;G) = H(G_{(Sp,Sp+n)}) - H(G_{(Sp,Sp+n)}|W_{(Sp,Sp+n)})$$

We compute MI over windows of eight consecutive samples (for higher contrast), shifted of one. Then we average this local MI measure over 500 consecutive samples, yielding an Average Mutual Information (AMI) measure for the position t in the sequence:

$$\mathrm{AMI}(S,t) = 1/500 \sum_{p=1,\ldots,500} (IM[S(t+p), S(t+p+8)](W;G))$$

Note that, for large n, two sequences having the same word frequencies have similar local MI. We represent in Fig. 7 AMI for the school book sequence SEQ, and for one of the RAND sequences. We see that AMI(RAND, t) is nearly constant. In contrast, AMI (SEQ, t) varies and its strongest variations are observed at times $t = 1$, 21, corresponding in the book sequence to some grouped word repetitions.

Interestingly, the ranks of the AMI and of the CWI are the same, we have: AMI(SORT) < AMI(SEQ) < AMI(RAND), and CWI(SORT) < CWI(SEQ) < CWI(RAND).

This suggests that the local mutual information criterion correlates well with word identification: words are better identified in a sequence that maximizes the mutual information of local stimuli. Further research will be conducted on this effect of sequence structure.

4.3. ART parameters and sequence order

Furthermore, our simulations revealed an interaction between the values of the learning parameters and presentation order on performance scores. When presentation order was based on order in the corpus, performance varied depending on V and L. In contrast, when presentation order was randomized, learning of infrequent words was improved.

These results suggest (1) that order of presentation is important especially for learning low frequency words, and (2) that L and V may mediate or control how ART can capitalize on this effect of order. We can further speculate on what parameters L and V might correspond to in children, but clearly they are interesting candidates to explain some of the variability observed in empirical data. Differences in these parameters could reflect individual differences and/or intra-individual variations in vigilance (arousal and attention) and readiness for learning. Future work could formally investigate these possibilities. This future research should also investigate to what extent the sensitivity of ART to order of presentation might be a good candidate for capturing empirical effects related to the age at which words are first learned (Age-of-Acquisition).

5. Conclusion

The ART network was shown to accurately reproduce the developmental pattern of word identification performance seen in children, for a specific combination of the learning and vigilance parameters in the model. The network reproduced the standard word frequency effect, and revealed a sensitivity to order of presentation of word stimuli. Word learning, particularly for low frequency words, was much improved when words were presented uniformly compared with the order in which they occur in reading textbooks. An analysis of the distribution of words in the reading corpus revealed a systematic grouping of low frequency words, having a smaller dispersion than high frequency words.

This raises the interesting possibility that order of presentation of words in current textbooks is suboptimal. From this point of view, ART might prove useful as a tool for evaluating school textbooks, by searching for optimal orders of presentation that would provide a benchmark for such evaluation. Finally, more attention will be dedicated to curriculum learning (Bengio et al., 2009) in the near future.

Acknowledgment

This work was supported by Grant (Agence Nationale de la Recherche, France) ANR-06-BLAN-0337 «Apprentissage neurocomputationnel de la lecture».

References

Balota, D., Chumbley, J., 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. Journal of Experimental Psychology: Human Perception & Performance 10 (3), 340–357.

Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada: ACM, pp. 41–48.

Carpenter, G.A., Grossberg, S., 1986a. Neural dynamics of category learning and recognition: attention, memory consolidation, and amnesia. In: Davis, J., Newburgh, R., Wegman, E. (Eds.), Brain Structure Learning and Memory. AAAS Symposium Series.

Carpenter, G.A., Grossberg, S., 1986b. Neural dynamics of category learning and recognition: structural invariants, reinforcement, and evoked potentials. In: Commons, M.L., Kosslyn, S.M., Herrnstein, R.J. (Eds.), Pattern Recognition and Concepts in Animals People, and Machines. Erlbaum, Hillsdale, NJ.

Carpenter, G.A., Grossberg, S., 1987. A massively parallel architecture for a self-organizing neural pattern recognition machine computer vision. Graphics and Image Processing 37, 54–115.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., Ziegler, J.C., 2001. DRC: a dual route cascaded model of visual word recognition and reading aloud. Psychological Review 108, 204–256.

Davis, C.J., 1999. The self-organising lexical acquisition and recognition (SOLAR) model of visual word recognition. Dissertation Abstracts International, 62, 594. <http://www.maccs.mq.edu.au/~colin/thesis.zip> (Doctoral dissertation, University of New South Wales, Sydney, New South Wales, Australia, 1999).

Dehaene, S., Cohen, L., Sigman, M., Vinckier, F., 2005. The neural code for written words: a proposal. Trends in Cognitive Sciences 9 (7), 335–341.

Diependaele, K., Ziegler, J., Grainger, J., in press. Fast phonology and the bi-modal interactive activation model. European Journal of Cognitive Psychology.

Dufau, S., Lété, B., Touzet, C., Glotin, H., Ziegler, J.C., Grainger, J., in press. A developmental perspective on visual word recognition: new evidence and a self-organizing model. European Journal of Cognitive Psychology.

Gomez, P., Ratcliff, R., Perea, M., 2008. The overlap model: a model of letter position coding. Psychological Review 115 (3), 577–600.

Goswami, U., Ziegler, J.C., 2006. A developmental perspective on the neural code for written words. Trends in Cognitive Sciences 10 (4), 142–143.

Grainger, J., 2008. Cracking the orthographic code: an introduction. Language and Cognitive Processes 23, 1–35.

Grainger, J., van Heuven, W., 2003. Modeling letter position coding in printed word perception. In: Bonin, P. (Ed.), The Mental Lexicon. Nova Science, Hauppause, NY, pp. 1–23.

Grainger, J., Granier, J.P., Farioli, F., Van Assche, E., van Heuven, W., 2006. Letter position information and printed word perception: the relative-position priming constraint. Journal of Experimental Psychology: Human Perception & Performance 32, 865–884.

Grainger, J., Rey, A., Dufau, S., 2008. Letter perception: from pixels to pandemonium! Trends in Cognitive Sciences 12, 381–387.

Grossberg, S., Boardman, I., Cohen, M., 1997. Neural dynamics of variable-rate speech categorization. Journal of Experimental Psychology: Human Perception & Performance 23 (2), 481–503.

Harm, M., Seidenberg, M., 1999. Phonology, reading acquisition, and dyslexia: insights from connectionist models. Psychological Review 106 (3), 491–528.

Harm, M., Seidenberg, M., 2004. Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. Psychological Review 111 (3), 662–720.

Hutzler, F., Ziegler, J.C., Perry, C., Wimmer, H., Zorzi, M., 2004. Do current connectionist learning models account for reading development in different languages? Cognition 91 (3), 273–296.

Lété, B., Sprenger-Charolles, L., Colé, P., 2004. Manulex: a grade-level lexical database from French elementary-school readers. Behavior Research Methods, Instruments, & Computers 36, 156–166.

McCloskey, M., Cohen, N.J., 1989. Catastrophic interference in connectionist networks: the sequential learning problem. In: Baa, G.H. (Ed.), The Psychology of Learning and Motivation, vol. 24. Academic Press, New York, pp. 109–165.

Perea, M., Lupker, S., 2004. Can CANISO activate CASINO! Transposed-letter similarity effects with nonadjacent letter positions. Journal of Memory and Language 51 (2), 231–246.

Perry, C., Ziegler, J.C., Zorzi, M., 2007. Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. Psychological Review 114, 273–315.

Plaut, D., McClelland, J., Seidenberg, M., Patterson, K., 1996. Understanding normal and impaired word reading: computational principles in quasi-regular domains. Psychological Review 103 (1), 56–115.

Rayner, K., Pollatsek, A., 1989. The Psychology of Reading. Lawrence Erlbaum Associates, Hillsdale, NJ.

Rubenstein, H., Garfield, L., Millikan, J., 1970. Homographic entries in the internal lexicon. Journal of Verbal Learning & Verbal Behavior 9 (5), 487–494.

Schoonbaert, S., Grainger, J., 2004. Letter position coding in printed word perception: effects of repeated and transposed letters. Language and Cognitive Processes 19 (3), 333–367.

Seidenberg, M., McClelland, J., 1989. A distributed, developmental model of word recognition and naming. Psychological Review 96 (4), 523–568.

Share, D.L., 1995. Phonological recoding and self-teaching: sine qua non of reading acquisition. Cognition 55 (2), 151–218.

Stevens, M., Grainger, J., 2003. Letter visibility and the viewing position effect in visual word recognition. Perception & Psychophysics 65, 133–151.

Whitney, C., 2001. How the brain encodes the order of letters in a printed word: the SERIOL model and selective literature review. Psychonomic Bulletin & Review 8, 221–243.

Ziegler, J.C., Goswami, U., 2006. Becoming literate in different languages: similar problems, different solutions. Developmental Science 9 (5), 429–436.

Ziegler, J.C., Perry, C., Coltheart, M., 2003. Speed of lexical and nonlexical processing in French: the case of the regularity effect. Psychonomic Bulletin & Review 10 (4), 947–953.

Zorzi, M., Houghton, G., Butterworth, B., 1998. Two routes or one in reading aloud? A connectionist dual-process model. Journal of Experimental Psychology: Human Perception & Performance 24 (4), 1131–1161.