LINGUE
E LINGUAGGIO

Anno X, N. 2, Dicembre 2011

SOMMARIO

A SELF-ORGANIZING MODEL OF WORD STORAGE AND PROCESSING: IMPLICATIONS FOR MORPHOLOGY LEARNING

MARCELLO FERRO CLAUDIA MARZI VITO PIRRELLI

ABSTRACT: In line with the classical cornerstone of "dual-route" models of word structure, assuming a sharp dissociation between memory and computation, word storage and processing have traditionally been modelled according to different computational paradigms. Even the most popular alternative to dual-route thinking − connectionist one-route models − challenged the lexicon-grammar dualism only by providing a neurally-inspired mirror image of classical base-to-inflection rules, while largely neglecting issues of lexical storage. Recent psycho- and neuro-linguistic evidence, however, supports a less deterministic and modular view of the interaction between stored word knowledge and on-line processing. We endorse here such a non modular view on morphology to offer a computer model supporting the hypothesis that they are both derivative of a common pool of principles for memory self-organization.

KEYWORDS: lexical processing, self organizing maps, morphological structure, serial memory.

1. INTRODUCTION

The mental lexicon is the store of words in long-term memory, where words are coded as time series of sounds/letters. From this perspective, the question of word coding, storage and maintenance in time is unseparable from the issue of how words are accessed and processed. In spite of this truism, lexical coding issues have suffered unjustified neglect by the Natural Language Processing and the Artificial Intelligence research communities. On the one hand, the unproblematic availability of primitive data structures such as ordered lists, strings, hierarchies and the like, recursively accessible through processing algorithms, has provided computer scientists with ready-made solutions to the problem of serial order representation. On the other hand, the mainstream connectionist answer to the problem of coding time series in artificial neural networks, so-called "conjunctive coding", appears to have eluded the problem rather than provide a principled solution.

In conjunctive coding (Coltheart *et al*., 2001; Harm & Seidenberg, 1999; McClelland & Rumelhart, 1981; Perry, Ziegler & Zorzi, 2007; Plaut *et al*., 1996), a word form like *cat* is represented through a set of context-sensitive episodic units. Each such unit ties a letter to a specific serial position (e.g. $\{C_1, A_2, T_3\}$), as in so-called positional coding or, alternatively, to a specific letter cluster (e.g. {_CA, CAT, AT_}), as customary in so-called Wickelcoding. Positional coding makes it difficult to generalize knowledge about phonemes or letters across positions (Plaut *et al*., 1996; Whitney, 2001) and to align positions across word forms of differing lengths (Davis & Bowers, 2004). The use of Wickelcoding, on the other hand, while avoiding some strictures of positional coding, raises the problem of the ontogenesis of representational units, which are hard-wired in the input layer. This causes an important acquisitional dead-lock. Speakers are known to exhibit differential sensitivity to symbol patterns. If such patterns are hard-wired in the input layer, the same processing architecture cannot be used to deal with languages exhibiting differential constraints on sounds or letters.

The failure to provide a principled solution to alignment issues is particularly critical from the perspective of morphology learning. Languages wildly differ in the way morphological information is sequentially encoded, ranging from suffixation to prefixation, sinaffixation, apophony, reduplication, interdigitation and combinations thereof. For example, alignment of lexical roots in three diverse pairs of paradigmatically related forms like English *walk-walked*, Arabic *kataba-yaktubu* 'he wrote'-'he writes' and German *machen-gemacht* 'make'-'made' (past participle) requires substantially different processing strategies. Coding any such strategy into lexical representations (e.g. through a fixed templatic structure separating the lexical root from other morphological markers) has the effect of slipping in morphological structure into the input, making input representations dependent on languages. A far more plausible solution would be to let the processing system home in on the right sort of alignment strategy through repeated exposure to a range of language-specific families of morphologically-related words. This is what conjunctive coding cannot do.

There have been three attempts to tackle the issue of time coding within connectionist architectures: Recursive Auto-Associative Memories (RAAM; Pollack, 1990), Simple Recurrent Networks (SRN; Botvinick & Plaut, 2006) and Sequence Encoders (Sibley *et al*., 2008). The three models set themselves different goals: i) encoding an explicitly assigned hierarchical structure for RAAM, ii) simulation of a range of behavioural facts of human Immediate Serial Recall for Botvinick & Plaut's SRNs and iii) long-term lexical entrenchment for the Sequence Encoder of Sibley and colleagues.

In spite of their differences, all systems model storage of symbolic sequences as the by-product of an auto-encoding task, whereby an input sequence of arbitrary length is eventually reproduced on the output layer after being internally encoded through recursive distributed patterns of node activation on the hidden layer(s). Serial representations and memory processes are thus modelled as being contingent on the task.

In this paper, we take a reversed approach to the problem. We describe a computational architecture for lexical storage based on Kohonen's Self-Organizing Maps (SOMs; Kohonen, 2001) augmented with first order associative connections encoding probabilistic expectations (so called Topological Temporal Hebbian SOMs, T2HSOMs for short; Koutnik, 2007; Pirrelli, Ferro & Calderone, in press; Ferro *et al*., 2010). The architecture mimics the behaviour of brain maps, medium to small aggregations of neurons in the cortical area of the brain, involved in selectively processing homogeneous classes of data. We show that T2HSOMs define an interesting class of general-purpose memories for serial order, exhibiting a non-trivial interplay between short-term and long-term memory processes. They simulate incremental processes of topological self-organization arranging lexical sequences in maximally predictive graphs and allow us to gain new insights into issues of grammar architecture and morphology learning.

2. BACKGROUND

According to the dual-route approach to word processing (Clahsen, 1999; Prasada & Pinker, 1993; Pinker & Prince, 1988; Pinker & Ullman, 2002), recognition of a morphologically complex input word involves two steps: i) preliminary full form access to the lexicon, ii) optional morpheme-based access of sub-word constituents, resulting from application of morphological rules of on-line word processing to the input word. Step ii) is taken if and only if step i) fails to find any matching entry in the lexicon. The approach endorses a direct functional correspondence between principles of grammar organization (lexicon *vs*. rules), processing correlates (storage *vs*. computation) and localization of the cortical areas functionally involved in word processing (temporo-parietal *vs*. frontal areas: Ullman, 2004).

Alternative theoretical models put forward a nuanced, indirect correspondence hypothesis, based on the emergence of morphological regularities from independent principles of organization of lexical information. In the Word-and-Paradigm tradition (Matthews, 1991; Pirrelli, 2000; Stump, 2001; Blevins, 2006), fully inflected forms are mutually related through possibly recursive paradigmatic structures, defining entailment relations between forms (Burzio, 2004). This view prompts a different computational metaphor than traditional rule-based models. A

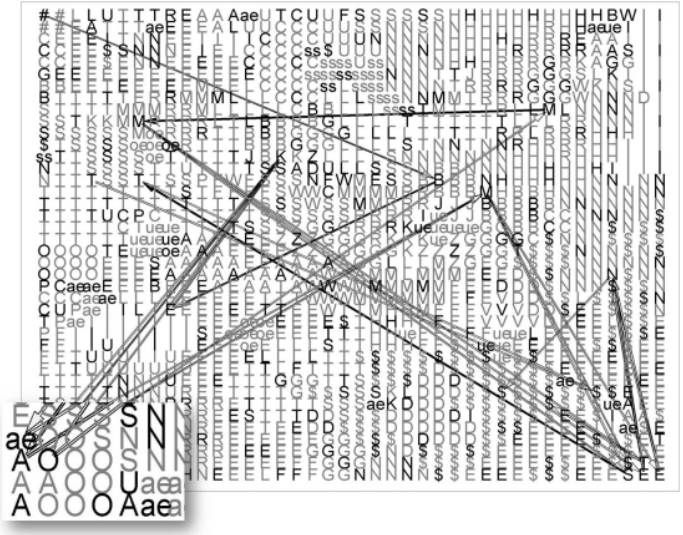
speaker's lexical knowledge corresponds more to one relational database, thus supporting a one-route model of word competence, than to a general-purpose automaton augmented with lexical storage (Blevins, 2006).

Over the past three decades, the psycholinguistic literature has shed novel light on this controversy. Recent empirical findings suggest that surface word relations constitute a fundamental domain of morphological competence, with particular emphasis on the interplay between form frequency, family frequency and family size effects within morphologically-based word families such as inflectional paradigms (Baayen, Dijkstra & Schreuder, 1997; Taft, 1979; Hay, 2001; Ford, Marslen-Wilson & Davis, 2003; Lüdeling & De Jong, 2002; Moscoso del Prado Fermìn *et al*., 2004; Stemberger & Middleton, 2003; Tabak, Schreuder & Baayen, 2005). However, that more than just lexical storage is involved is suggested by interference effects between false morphological friends (or pseudo-derivations) such as *broth* and *brother*, sharing a conspicuous word onset but unrelated morphologically (Frost, Forster & Deutsch, 1997; Rastle, Davis & New, 2004; Post *et al*., 2008). The evidence shows that as soon as a given letter sequence is fully decomposable into morphological formatives, word parsing takes place automatically, prior to (or concurrently with) lexical look-up. The emerging view sees word processing as the outcome of simultaneously activating patterns of cortical connectivity reflecting redundant distributional regularities in input data at the phonological, morpho-syntactic and morpho-semantic levels. This suggests that differentiated brain areas devoted to language maximize the opportunity of using both general and specific information simultaneously (Libben, 2006; Post *et al*., 2008), rather than maximize processing efficiency and economy of storage.

T2HSOMs adhere to such a dynamic, non modular view of the interaction between memory and computation, whereby word processing and learning are primarily conceived of as memory-driven processes. They part from both dual-route and one-route approaches in supporting the view that the way words are structured in our long-term memory is key to understanding the mechanisms governing word processing. This perspective focuses on word productivity as the by-product of more basic memory processes that must independently be assumed to account for word learning. Secondly, it opens up new promising avenues of inquiry by tapping the large body of literature on short-term and long-term memories for serial order (see Baddley, 2007, for an overview). Furthermore, it gives the opportunity of using sophisticated computational models of language-independent memory processes (Botvinick & Plaut, 2006; Brown, Preece & Hulme, 2000; Burgess & Hitch, 1996, among others) to shed light on language-specific aspects of word encoding and storage.

3. TOPOLOGICAL TEMPORAL SOMs

T2HSOMs are grids of topologically organized memory nodes, exhibiting dedicated sensitivity to time-bound stimuli. Upon presentation of an input stimulus, all map nodes are activated synchronously, but only the most sensitive node to the incoming stimulus, the so-called Best Matching Unit (BMU), wins over the others. Figure 1 illustrates the chains of BMUs triggered by 9 forms of German BEKOMMEN 'become' on a 40x40 nodes map: *bekam* 'became' (1S/3S past tense), *bekäme* 'became' (1S/3S past subj), *bekamen* 'became' (1P/3P past tense), *bekämen* 'became' (1P/3P past subj), *bekomme* 'become' (1S pres ind, 1S pres subj), *bekommen* 'become' (inf, 1P/3P pres ind, past participle, 1P/3P pres subj), *bekommst* 'become' (2S pres ind), *bekommt* 'becomes' (3S pres ind), *bekämst* 'became' (2S past subj). The map was trained on 103 verb paradigms, sampled from the Celex German database, for a total amount of 881 verb form types with different frequency distributions.



Trained forms:
bekam
bekäme
bekamen
bekämen
bekomme
bekommen
bekommst
bekommt

Untrained form:
bekämst

FIGURE 1. ACTIVATION CHAINS FOR 9 INFLECTED FORMS OF GERMAN *BEKOMMEN* 'BECOME'. UMLAUTED VOWELS ARE CODED AS DIPHTHONGS ON THE MAP.

In Figure 1, each node is labelled with the letter the node is most sensitive to.[1] Note that letters are encoded using orthogonal vectors (localist coding) so that any symbol is equally distant from any other one.[2] Pointed

[1] Each node is assigned a letter label as a thresholded function of $y_{S,i}(t)$, according to equation (2) below (see section 5.1 for more detail).

[2] The model is a sequence encoder and is agnostic as to nature and type (e.g. localist *vs*. distributed) of input representations. Reported experiments make use of input data consisting of orthographic transcriptions; results, however, are not contingent upon the specific characteristics of a language orthography, and the same model can be applied to any input

arrows depict the temporal sequence of node exposure (and node activation), starting from a beginning-of-the-word symbol # (anchored in the top left corner of the map) and ending with the end-of-the-word symbol $. The thickness of arrows represents the strength of the corresponding temporal connections. The magnified bottom left corner of the figure corresponds to an area of nodes that show sensitivity to stem ablauting (as in *bekommen*, *bekam*, *bekäme*). As will be clearer from the ensuing sections, the topological proximity of alternating vowels in stem allomorphs is the result of their being systematically distributed in (nearly) identical contexts. In turn, topological proximity favours i) convergence of the corresponding activation chains and ii) the emergence of a notion of abstract stem as a receptive field of the map.

3.1 The architecture

Dedicated sensitivity and topological organization are not hard-wired on the map but are the result of self-organization through learning, whereby neighbouring nodes get increasingly sensitive to input symbols (letters) that are similar in both encoding and distribution.

Figure 2 offers an overview of the architecture of a T2HSOM. Map nodes present two levels of connectivity. First, they are fully connected with the input vector through connections with no time delay, forming the spatial connection layer. Weights on spatial connections are adjusted during learning to better respond to input stimuli. Secondly, nodes are mutually connected through a temporal layer, whose connections are updated with a fixed one-step time delay, based on activity synchronization between BMUs.

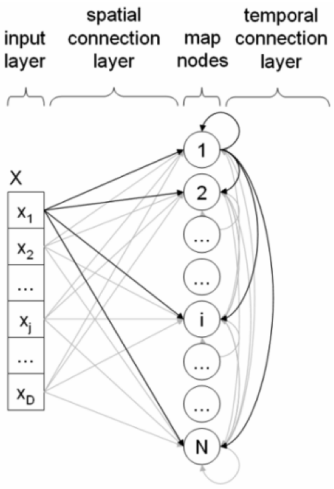


FIGURE 2. OUTLINE ARCHITECTURE OF A T2HSOM

transcription based on symbol concatenation or phonological coding.

Each learning step includes three phases: input encoding, activation and weight adjustment. A symbol is represented on the input layer at time *t* through an input vector of *x* codes. At each exposure, map nodes are activated in parallel as a function of i) how close their spatial connection weights are to *x* codes of the current input vector, and ii) how strongly nodes are synaptically connected with the BMU at time *t-1* over the temporal layer. More formally, the activation $y_i(t)$ of the *i*-th node of the map at time *t* is:

$$(1) \qquad y_i(t) = \alpha \cdot y_{S,i}(t) + \beta \cdot y_{T,i}(t)$$

In equation (1), α and β weigh up the respective contribution of the spatial ($y_{S,i}(t)$) and temporal layer ($y_{T,i}(t)$) to the overall activation level. $y_{S,i}(t)$ is calculated on the basis of code similarity, as classically modelled by Kohonen's SOMs (Kohonen, 2001). Each node is activated as an inverse function of the distance between the node's vector of synaptic weights on the spatial connection layer and the current input vector. Using the Euclidean distance to measure code similarity, the contribution of the *i*-th node on the spatial layer at time *t* is:

$$(2) \qquad y_{S,i}(t) = \sqrt{D} - \sqrt{\sum_{j=1}^{D} [x_j(t) - w_{i,j}(t)]^2}$$

where $x(t)$ is the *D*-dimensional input vector and $w_i(t)$ is the *D*-dimensional spatial weight vector of the *i*-th node. On the other hand, the contribution on the temporal layer is calculated on the basis of a synchronization principle. According to the Hebbian rule, the synapses between two neurons get stronger if the neurons show a tendency to fire at a short time distance, and they get weaker if the neurons normally do not fire at short time intervals. Using a dot product to evaluate activity synchronization, the contribution of the *i*-th node on the temporal layer at time *t* is:

$$(3) \qquad y_{T,i}(t) = \sum_{h=1}^{N} [y_h(t-1) \cdot m_{i,h}(t)]$$

representing the weighted temporal pre-activation of the *i*-th node at time *t* prompted by the state of activation of all *N* nodes at time *t-1* (namely $y(t\text{-}1)$) and the *N*-dimensional temporal weight vector of the *i*-th node (namely $m_i(t)$). As a result, weight adjustment affects i) weights on the spatial layer for them to get closer to the corresponding values on the input layer and ii) weights on the temporal layer for them to synchronize with previously activated BMUs.

Weight adjustment does not apply evenly across map nodes and learning epochs, but is a function of the map's learning rate and space topology. At

each activation step, the current BMU is adjusted most strongly, while all other nodes get adjusted as a Gaussian function of their distance from the BMU (or neighbourhood function). The learning rate defines how quickly weights are adjusted at each learning epoch, simulating the behaviour of a brain map adapting its plasticity through learning. More formally, plasticity of the spatial connection weights is driven by the similarity error:

$$\Delta w_{i,j}(t) \propto [x_j(t) - w_{i,j}(t)] \tag{4}$$

and plasticity of the temporal connection weights is driven by the synchronization error:

$$\Delta m_{i,h}(t) \propto \begin{cases} 1 - m_{i,h}(t) & potentiation \\ m_{i,h}(t) & depression \end{cases} \tag{5}$$

Figure 3 pictorially illustrates the relationship between equation (5) and the neighbourhood function on the temporal connection layer. Unlike classical conjunctive representations in either Simple Recurrent Networks (Elman, 1990) or Recursive SOMs (Voegtlin, 2002), where both order and item information is collapsed on the same layer of connectivity, T2HSOMs keep the two types of information stored on separate (spatial and temporal) layers, which are trained according to independent principles, namely code similarity and firing synchronization.

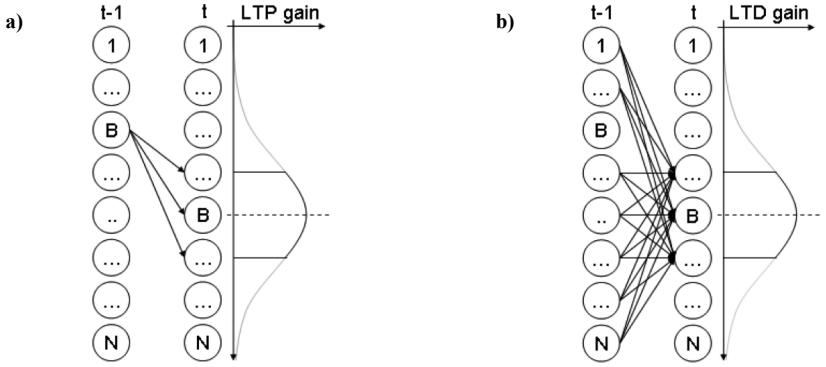


FIGURE 3. WEIGHT ADJUSTMENT AND NEIGHBOURHOOD FUNCTION ON THE TEMPORAL LAYER OF A T2HSOM.

3.2 Memory structures and memory orders

By being repeatedly exposed to word forms encoded as temporal sequences of letters, a T2HSOM tends to dynamically store strings through a graph-like hierarchical structure of nodes. A graph starts with a # node and branches

out when different nodes are alternative continuations of the same history of activated nodes (Figure 1). The length of the history of past activations defines the order of memory of the map. It can be shown that this type of organization maximizes the map's expectation of an upcoming symbol in the input string or, equivalently, minimizes the entropy over the set of transition probabilities from one BMU to the ensuing one. This prompts a process of incremental specialization of memory resources, whereby several nodes are recruited to be sensitive to contextually specific occurrences of the same letter.

The ability to store a word form through a uniquely dedicated chain of BMUs depends on the order of memory of a T2HSOM. It can be shown that the order of memory of a T2HSOM is, in turn, a function of i) the size of the map (i.e. the number of nodes), and ii) the map's ability to train two adjacent nodes independently.

Figure 4 illustrates how this process of incremental specialization unfolds through training. For simplicity, we are assuming a map trained on two strings only: #a1 and #b1. Figure 4a represents an early stage of learning, when the map recruits a single BMU for the symbol *1* irrespective of its embedding context. After some learning epochs, two different BMUs are concurrently activated for *1* through equally strong connections (Figure 4b). Connections get increasingly specialized in Figure 4c, where the two *1* nodes are preferentially selected upon seeing either *a* or *b*. Finally, Figure 4d illustrates a stage of dedicated connections, where each *1* node is selected by one specific left context only. The stage is reached when the map can train each single node without affecting any neighbouring node. Technically, this corresponds to a learning stage where the map's neighbourhood radius is *0*.

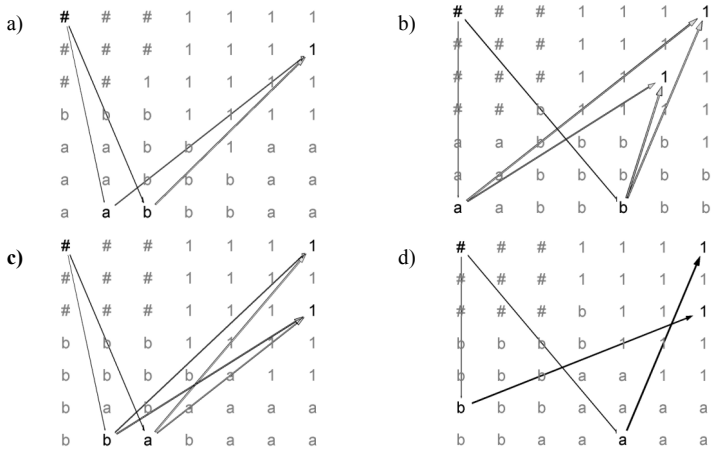


FIGURE 4. STAGES OF CHAIN DEDICATION, FROM EARLY TO FINAL LEARNING EPOCHS.

4. LEXICAL PROCESSING AND STORAGE

T2HSOMs are models of both string processing and string storage. In activation mode, they provide an incremental coding for incoming input signals. For each input symbol, the map's code for that symbol is the weight vector on the spatial connection layer associated with the corresponding BMU (Figure 2 above). We can therefore estimate the map's accuracy in processing an input symbol at time *t* as a function of $y_{S,BMU}(t)$, according to equation (2). We can also monitor the map's capacity to memorize strings by simulating a test of lexical recall. Lexical recall is modelled here as the task of generating a word form *w* from the integrated pattern of node activation triggered by *w*. The task is coherent with Baddeley's view of the interaction between the short-term and the long-term memory stores in the speaker's brain. When the map is exposed to a sequence of symbols, the activation pattern triggered by each symbol is rehearsed in a short-term buffer. As more patterns (one for each symbol in the input sequence) are rehearsed simultaneously, the resulting activation state of the short-term buffer is the integration of more overlaying patterns (see Figure 5). Lexical recall consists in feeding the (long-term) lexical store (a trained map) with such an integrated short-term pattern to see if the former can generate all input symbols in the appropriate order. Since all symbols are presented simultaneously, this is possible only if the lexical map has developed appropriate temporal expectations on incoming input symbols.

More formally, we define the integrated activation pattern $\hat{Y}=\{\hat{y}_1,\ldots,\hat{y}_N\}$ of a word of *n* symbols as the result of choosing:

$$\hat{y}_i = \max_{t=2,\ldots,n}\{y_i(t)\} \qquad i=1,\ldots,N \tag{6}$$

Lexical recall is thus modelled by the activation function of equation (1) above, with:

$$y_{S,i}(t) = \sqrt{D} - \sqrt{\sum_{j=1}^{D}[x_j(t) - w_{i,j}]^2} \tag{7}$$

for *t=1* (i.e. when the map is primed with #), and:

$$y_{S,i}(t) = \hat{y}_i(t) \tag{8}$$

for *t=2,...,n*.

This is a considerably more difficult task than activating a specific node upon seeing a particular input symbol at time *t*. For a map to be able to correctly reinstate a whole string *s* from its integrated short-term pattern, a time-bound activation chain dedicated to *s* must be memorized in the

long-term store. This means that the map has to develop, through learning, a strong expectation to see the incoming input. The strength of such predictive drive is measured by $y_{T,i}(t)$ in equation (3) above. Lexical recall probes this long-term time-bound expectation.

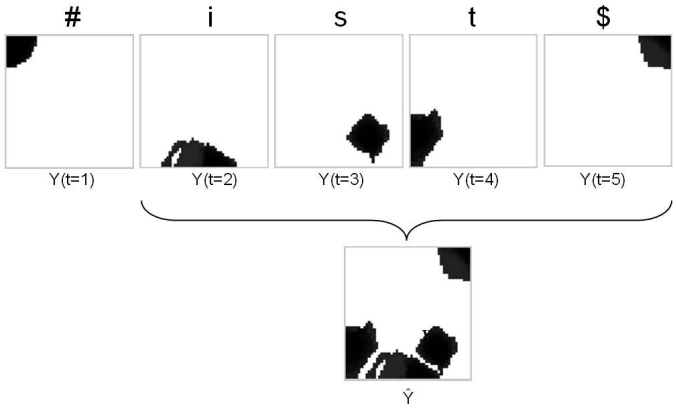


FIGURE 5. PER-LETTER AND INTEGRATED SHORT-TERM ACTIVATION PATTERN FOR #IST$.

5. EXPERIMENTAL EVIDENCE

5.1 Experiment 1

We tested the dynamic behaviour of two 40x40 maps on two tasks: activation and lexical recall. One map was trained on 1672 Italian inflected verb types (3901 tokens) sampled from the Italian TreeBank (Montemagni *et al*., 2003). We trained the second map on 881 German inflected verb types (4995 tokens) from the German section of the Celex database (Baayen, Piepenbrock & Gulikers, 1996), based on the Manheim corpus frequency distributions. For both languages, low frequency verb forms that were not used for training but were part of verb paradigms shown in training were set aside to form a test set of novel (untrained) words (see infra).

Each word form was input to the map as a letter string preceded by # and followed by $ (e.g. #IST$ for *ist* 'is'), with all letters common to the Italian and German alphabets written in upper-case. Umlauted characters were written as lower-case diphthongs (e.g. #BEKaeME for *bekäme* 'became') and the sharp s *ß* as *ss* (e.g. #HEIssEN$ for *heißen* 'call').[3] On the input layer, letters were recoded as mutually orthogonal binary vector codes (localist coding). Identical letter codes were used for upper-case

[3] In both cases, pairs of lower-case letters were processed as forming a single orthographic symbol.

letters in Italian and German. Both maps were trained on 100 epochs, with α=0.1 and β=1. At each epoch all training forms were shown to the map, one letter at a time, with the map temporal expectation being reset upon presentation of the $ symbol. Each word was randomly shown to the map according to its probability distribution in the source corpus. Accordingly, more frequent words were presented to the map more often than less frequent words.

We first assessed how well each map could process and recall verb forms that were part of the map's training set (see Table 1). In the activation task, we estimated the map's accuracy in processing an input symbol at time t as a function of $y_{S,BMU}(t)$, where $y_{S,BMU}(t)$ is obtained from equation (2) above with $i = BMU$. In particular, the map is taken to process the current input letter accurately if $\sqrt{D} - y_{S,BMU}(t) < 0.1$. The same inequality is also used to estimate the map's accuracy in recalling the input symbol # at time $t=1$. For t=2,…,n, we use the inequality $\sqrt{D} - \hat{y}_{S,BMU}(t) < 0.1$, where $\hat{y}_{S,BMU}(t)$ is obtained from equation (8) above with $i = BMU$.

We also tested the map's response to a set of untrained verb forms (or test set) belonging to verb paradigms shown in training. The Italian test set contained 484 verb forms, the German test set contained 188 verb forms. Table 1 also gives the results of probing the Italian map on the German test set (Italian non-words) and, conversely, probing the German map on the Italian test set (German non-words). This was done on both tasks. Overall results are reported in terms of percentage of per-word accuracy: each input word is taken to be processed or recalled accurately if all its symbols are processed or recalled accurately.

		% ACCURACY	
		Italian	German
PROCESSING	training set	100	100
	test set	97.9	96.2
	non-words	53.7	52.9
LEX RECALL	training set	91.0	99.6
	test set	82.4	81.9
	non-words	3.7	3.1

TABLE 1

5.2 Experiment 2

To test the map's capacity of developing expectations on the morphological structure of the trained verb forms, we used a modified version of Albright's (2002) experimental protocol. We selected 34 novel target forms: 17 Italian infinitives (e.g. #SEMBRARE$ 'to seem') and 17 Italian second person plural present indicative forms (e.g. #SEMBRATE$ 'you seem') that were not

shown in training but were part of the inflectional paradigms the Italian map was trained on. In Italian, both the infinitive and second present indicative forms contain one of three possible thematic vowels (*a*, *e* or *i*), depending on their conjugation class. For each target form, we added to the test set two nonce forms, generated by replacing the appropriate thematic vowel with the two other vowels. For example, if the target form is #SEMBRARE$, we included the two nonce forms #SEMBRERE$ and #SEMBRIRE$. Testing the map's response on this set allowed us to assess how well the map recalls the correct form and its ungrammatical competitors. The expectation is that if the map develops a sensitivity to the paradigmatic structure of Italian conjugation, it should be able to assign higher scores to a correct unseen verb form, based on training evidence. Here the word score *S* for a word of *n* symbols is a function of the map's ability to match the input symbol (S_S) and to predict it (S_T):

$$S = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{2}S_S(t) + \frac{1}{n-1}\sum_{t=2}^{n}\frac{1}{2}S_T(t) \tag{9}$$

with

$$S_S(t) = \sqrt{D} - \sqrt{\sum_{j=1}^{D}\left[x_j(t) - w_{BMU(t)\,j}\right]^2} \tag{10}$$

with $t=1,\ldots,n$ and

$$S_T(t) = m_{BMU(t)\,BMU(t-1)} \tag{11}$$

with $t=2,\ldots,n$. Overall results are shown in the top half of Table 2 below. Note that the difference between the average recall scores on correct (0.475) and nonce verb forms (0.409) is statistically significant (*p-value* < 0.01).

	CORRECT FORMS	MADE UP FORMS
Count	34	68
Processing accuracy	100%	100%
Recall accuracy	64.7%	35.3%
Average processing score	0.543	0.542
Average recall score	0.475	0.409
Processing hits	50%	
Recall hits	67,6%	

TABLE 2

We can also conceptualize our test as a classification task. For each verb triple (e.g. #SEMBRERE$, #SEMBRERE$ and #SEMBRIRE$) the map classifies the form with the highest *S* score as the morphologically

correct form. Once more, the score was calculated in both processing and recall. Figures are reported in the bottom half of Table 2, in terms of the percentage of hits (number of correct responses) out of all tested triples.

6. DISCUSSION AND CONCLUDING REMARKS

T2HSOMs exhibit a remarkable capacity of recoding an incoming word form correctly, through activation of the contextually appropriate BMUs. This is a very robust behaviour, as shown by our experimental evidence, mostly stemming from accurate spatial recoding of input letters. That also expectations are involved, however, is shown by the processing errors on non-words in our first experiment (Table 1). Here, the Italian map was tested on German forms and the German map was tested on Italian forms. Results bear witness to the discrepancy between acquired expectations and unexpectedly novel evidence.

Processing expectations, however, are fairly local, contrary to recall expectations which can hold over longer stretches of letters. By definition, recall is based on the map's capacity of anticipating upcoming symbols based on an acquired predictive drive. This is what dynamic storage is about.

Storage involves recruitment of dedicated memory chains, i.e. chains of context-sensitive nodes keeping track of repeatedly seen sequences of letters. The map's sensitivity to frequency in recall is shown by the higher recall rate on training German verb forms, which present higher frequency distributions than the corresponding Italian forms (Table 1). Dedicated chains take map's space, as they require recruitment of specialized context-sensitive nodes which fire only when a particular symbol is presented within a particular word or word family. Once more, the German training set, with fewer word types, makes the map more proficient in recalling familiar word forms. As a general remark, paradigmatic families presenting radically suppletive forms (e.g. *go-went*), unlike morphologically regular paradigms, are not conducive to chain sharing. Cases of lexically-conditioned morphological alternation like *bring-brought*, on the other hand, will be subject to an intermediate storage strategy, between dedicated and shared chains, the amount of shared morphological structure depending on both token and type frequency of the specific morphological alternation.

A further important aspect of dynamic storage has to do with generalization. Not only is the map in a position to access and recall familiar strings, but it can also build up expectations about novel combinations of letters. The map structures redundant information through shared activation chains, thus making provision for chain combinations that are never triggered in the course of training. The effect is reminiscent of what is illustrated in Figure 4 above, where wider neighbourhoods, typical

of early stages of learning, favour more liberal inter-node connections. In experiment 1, the map is too small to be able to dedicate a different node to a different context-dependent occurrence of a letter. Fewer nodes are recruited to be sensitive to several different context-sensitive tokens of the same letter type and to be more densely connected with other nodes. A direct consequence of this situation is generalization, corresponding to the configurations shown in Figure 4b and 4c above. Most notably, this is the by-product of the way the map stores and structures lexical information.

Experiment 2 throws in sharp relief a further important issue: memory expectations are sensitive to morphological structure. Note that recall accuracy – i.e. the map's capacity of reinstating a novel word form – is a direct function of the form's morphological coherence. Furthermore, the average recall score on paradigmatically coherent forms is significantly higher that the corresponding score on paradigmatically spurious forms. Such a difference is statistically not significant in the processing score, which, once more, proves to be sensitive to more local constraints.

It is traditionally assumed that structure-driven generalizations take centre stage in language learning, playing the role of default on-line mechanisms in language processing. From this perspective, the lexicon is a fall back, costly store of exceptions, conveying a comparatively minor portion of language competence. The evidence presented here shows that another view is possible. Word processing and learning are primarily memory-driven processes. Pre-compilation of dedicated long-term memory chains is beneficial for prediction in on-line word processing. Morpheme-based generalizations, on the other hand, represent shorter chains that come into the picture when memory of longer chains (whole words) fails, due to either novel, degenerate and noisy input, or to limitations in perception/memory spans. This explains the remarkably conservative nature of language learning, where over-regularization and levelling effects take place occasionally, and supports a more dynamic and less modularized view of language processing, where memory and computation, holistic and combinatorial knowledge, are possibly two sides of the same coin.

REFERENCES

Albright, A. (2002). Islands of reliability for regular morphology: evidence from Italian. *Language* 78, 684-709.

Baayen, H., Dijkstra, T. & Schreuder, R. (1997). Singulars and plurals in Dutch: evidence for a parallel dual route model. *Journal of Memory and Language* 37, 94-117.

Baayen, H., Piepenbrock, R. & Gulikers, L. (1996). *CELEX*. Philadelphia: Linguistic Data Consortium.

Baddeley, A. D. (2007). *Working Memory, Thought and Action*. Oxford: Oxford University Press.

Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics* 42, 531-573.

Botvinick, M. & Plaut, D. C. (2006). Short-term memory for serial order: a recurrent neural network model. *Psychological Review* 113, 201-233.

Brown, G. D. A., Preece, T. & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review* 107, 127-181.

Burgess, N. & Hitch, G. J. (1996). *A connectionist model of STM for serial order*. In S. E. Gathercole (Ed.), *Models of short-term memory* (pp. 51-71). Hove: Psychology Press.

Burzio, L. (2004). Paradigmatic and syntagmatic relations in italian verbal inflection. In J. Auger, J. C. Clements & B. Vance (Eds.), *Contemporary Approaches to Romance Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Clahsen, H. (1999), Lexical entries and rules of language: a multidisciplinary study of German inflection. *Behavioral and Brain Sciences* 22, 991–1060.

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

Davis, C. J. & Bowers, J. S. (2004). What do letter migration errors reveal about letter position coding in visual word recognition? *Journal of Experimental Psychology: Human Perception and Performance* 30, 923-941.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science* 14 (2), 179-211.

Ferro, M., Ognibene, D., Pezzulo, G. & Pirrelli, V. (2010). Reading as active sensing: a computational model of gaze planning in word recognition. *Frontiers in Neurorobotics* 4, 1-16.

Ford, M., Marslen-Wilson, W. & Davis, M. (2003). Morphology and frequency: contrasting methodologies. In H. Baayen & R. Schreuder (Eds.), *Morphological Structure in Language Processing*. Berlin/New York: Mouton de Gruyter.

Frost, R., Forster, K. I. & Deutsch, A. (1997). What can we learn from the morphology of Hebrew? A masked priming investigation of morphological re-presentation. *Journal of Experimental Psychology: Learning, Memory and Cognition* 23, 829-856.

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

Hay, J. (2001). Lexical frequency in morphology: is everything relative? *Linguistics* 39, 1041-1111.

Kohonen, T. (2001). *Self-Organizing Maps*. Heidelberg: Springer-Verlag.

Koutnik, J. (2007). Inductive modelling of temporal sequences by means of self-organization. *Proceeding of International Workshop on Inductive Modelling (IWIM 2007)* (pp. 269-277). Prague.

Libben, G. (2006). Why studying compound processing? An overview of the issues. In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words* (pp. 1-22). Oxford: Oxford University Press.

Lüdeling, A. & Jong, N. de (2002). German particle verbs and word formation. In

N. Dehé, R. Jackendoff, A. McIntyre & S. Urban (Eds.), *Explorations in Verb-Particle Constructions* (pp. 315-333). Berlin/New York: Mouton de Gruyter.

Matthews, P. H. (1991). *Morphology*. Cambridge: Cambridge University Press.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review* 88, 375-407.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M. T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F. & Delmonte, R. (2003). Building the Italian syntactic-semantic treebank. In A. Abeillè (Ed.), *Building and Using Parsed Corpora* (pp. 189-210). Dordrecht: Kluwer.

Moscoso del Prado Fermìn, M., Bertram, R., Häikiö, T., Schreuder, R. & Baayen, H. (2004). Morphological family size in a morphologically rich language: the case of Finnish compared with Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition* 30 (6), 1271-1278.

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 (2), 273-315.

Pinker, S. & Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* 29, 195-247.

Pinker, S. & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Science* 6, 456-463.

Pirrelli, V. (2000). *Paradigmi in morfologia. Un approccio interdisciplinare alla flessione verbale dell'italiano*. Pisa/Roma: Istituti Editoriali e Poligrafici Internazionali.

Pirrelli V., Ferro, M. & Calderone, B. (in press). Learning paradigms in time and space. Computational evidence from Romance languages. In M. Goldbach, M.-O. Hinzelin, M. Maiden & J. C. Smith (Eds.), *Morphological Autonomy: Perspectives from Romance Inflectional Morphology*. Oxford: Oxford University Press.

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

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence* 46, 77–105.

Post, B., Marslen-Wilson, W., Randall, B. & Tyler, L. K. (2008). The processing of English regular inflections: phonological cues to morphological structure. *Cognition* 109, 1-17.

Prasada, S. & Pinker, S. (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes* 8, 1-56.

Rastle, K., Davis M. H. & New, B. (2004). The broth in my brother's brothel: morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin and Review* 11 (6), 1090-1098.

Sibley, D. E., Kello, C. T., Plaut, D. & Elman, J. L. (2008). Large-scale modeling of wordform learning and representation. *Cognitive Science* 32, 741-754.

Stemberger, J. P. & Middleton, C. S. (2003). Vowel dominance and morphological processing. *Language and Cognitive Processes* 18 (4), 369-404.
Stump, G. T. (2001). *Inflectional morphology*. Cambridge: Cambridge University Press.
Tabak, W., Schreuder, R. & Baayen, R. H. (2005). Lexical statistics and lexical processing: semantic density, information complexity, sex and irregularity in Dutch. In M. Reis & S. Kepser (Eds.), *Linguistic Evidence* (pp. 529-555). Berlin/New York: Mouton de Gruyter.
Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory and Cognition* 7, 263-272.
Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition* 92, 231-270.
Voegtlin, T. (2002). Recursive self-organizing maps. *Neural Networks* 15, 979–991.
Whitney, C. (2001). How the brain encodes the order of letters in a printed word: the SERIOL model and selective literature review. *Psychonomic Bulletin and Review* 8, 221-243.

Marcello Ferro
Istituto di Linguistica Computazionale "A. Zampolli"
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1, 56124 Pisa
Italy
e-mail: marcello.ferro@ilc.cnr.it

Claudia Marzi
Istituto di Linguistica Computazionale "A. Zampolli"
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1, 56124 Pisa
Italy
e-mail: claudia.marzi@ilc.cnr.it

Vito Pirrelli
Istituto di Linguistica Computazionale "A. Zampolli"
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1, 56124 Pisa
Italy
e-mail: vito.pirrelli@ilc.cnr.it