T2HSOM: Understanding the Lexicon by Simulating Memory Processes for Serial Order

Marcello Ferro, Claudia Marzi, Vito Pirrelli
Institute for Computational Linguistics "A. Zampolli", National Research Council
via G. Moruzzi 1, Pisa, Italy
e-mail: {marcello.ferro, claudia.marzi, vito.pirrelli}@ilc.cnr.it

Abstract

Over the last several years, both theoretical and empirical approaches to lexical knowledge and encoding have prompted a radical reappraisal of the traditional dichotomy between lexicon and grammar. The lexicon is not simply a large waste basket of exceptions and sub-regularities, but a dynamic, possibly redundant repository of linguistic knowledge whose principles of relational organization are the driving force of productive generalizations. In this paper, we overview a few models of dynamic lexical organization based on neural network architectures that are purported to meet this challenging view. In particular, we illustrate a novel family of Kohonen self-organizing maps (T2HSOMs) that have the potential of simulating competitive storage of symbolic time series while exhibiting interesting properties of morphological organization and generalization. The model, tested on training samples of as morphologically diverse languages as Italian, German and Arabic, shows sensitivity to manifold types of morphological structure and can be used to bootstrap morphological knowledge in an unsupervised way.

1. Introduction

Traditional generative approaches to language inquiry view word competence as consisting of a morphological lexicon, an assorted hotchpotch of exceptions and sub-regularities, and a grammar, a set of productive combinatorial rules (Di Sciullo and Williams 1987; Prasada and Pinker 1993). Whatever cannot be assembled through rules must be relegated wholesale to the lexicon, whose size depends on the generative power of the grammar: the richer the power, the poorer the lexicon.

Baayen (2007) observes that the approach reflects an outdated view of lexical storage as more 'costly' than computational operations. Similarly, alternative theoretical models question the primacy of grammar rules over lexical storage, arguing that morphological regularities emerge from independent principles of lexical organization, whereby fully inflected forms are redundantly stored and mutually related through entailment lexical relations (Matthews 1991; Pirrelli 2000; Burzio 2004; Blevins 2006). This view prompts a radically different computational metaphor than traditional generative models. A speaker's knowledge corresponds more to one large dynamic relational database than to a general-purpose automaton augmented with lexical storage.

In spite of the large body of theoretical literature on the topic, however, few computational models of the lexicon can be said to address such a complex interaction between storage and computation. Contrary to what is commonly held, connectionism has failed to offer an alternative view of the interplay between lexicon and grammar. As we shall argue in more detail in the ensuing session, there is no place for the lexicon in classical connectionist networks. Somewhat ironically, they seem to have adhered to a cornerstone of the rule-based approach to morphological inflection, thus providing a neurally-inspired mirror image of inflection rules.

In this paper, we will explore the somewhat complementary view that storage plays a fundamental role in lexical modelling, and that computer simulations of short-term and long-term memory processes can go a long way in addressing issues of lexical organization. The present paper lends support to this claim by illustrating a novel neural network architecture known as "Topological Temporal Hebbian Self-Organizing Map" (or T2HSOM for short, Ferro *et al.* 2010). A T2HSOM has the potential of simulating dynamic storage of symbolic time series while exhibiting interesting properties of morphological self-organization. Trained on morphologically diverse families of word forms, T2HSOMs can be shown to bootstrap morphological structure in an unsupervised way. Finally, we suggest that they offer an ideal workbench for understanding the structure of the lexicon by simulating memory processes.

2. Background

As a first approximation, the lexicon is the store of words in long-term memory. Any attempt at modelling lexical competence must hence take issues of string storage very seriously. In this respect, the rich cognitive literature on short-term and long-term memory processes (Miller 1956; Baddeley and Hitch 1974; Baddeley 1986; 2006; Henson 1998; Cowan 2001; among others) has the unquestionable merit of highlighting some fundamental issues of coding, maintenance and manipulation of time-bound constraints over strings of symbols.

Word forms are primarily sequences of sounds or letters and so the question of their coding (and maintenance) in time is logically prior to any other processing issue. In spite of this truism, however, coding issues have suffered unjustified neglect by the NLP research community over the last 30 years. In fact, the mainstream connectionist answer to the problem of time series coding, namely so-called "conjunctive coding", appears to elude some core issues in lexical representation.

Conjunctive codes (e.g., Coltheart, Rastle, Perry, Langdon and Ziegler 2001; Harm and Seidenberg 1999; McClelland and Rumelhart 1981; Perry, Ziegler, and

Zorzi 2007; Plaut, McClelland, Seidenberg, and Patterson 1996) are typically assumed to be available in the input (or encoding) layer of a multi-layered perceptron in the form of a built-in repertoire of context-sensitive Wickelphones, such as $_{\#}C_{a}$ and $_{c}A_{t}$ to respectively encode the letters *c* and *a* in *cat*. However, the use of Wickelphones raises the immediate issue of their ontogenesis, since they appear to solve the problem of coding time series by resorting to time-bound relations whose representation in the encoding layer remain unexplained. A second related issue is the acquisition of phonotactic knowledge. Speakers are known to exhibit differential sensitivity to diverse sound patterns. Effects of graded specialization in the discrimination of sound clusters and lexical well-formedness judgements are the typical outcome of acquiring a particular language. If such patterns are part and parcel of the encoding layer, the same processing system cannot be used to deal with different languages exhibiting differential sound constraints.

A third limitation of conjunctive coding is that phonemes and letters are bound with their context. This means that two elements like $_{\#}E_{v}$ and $_{v}E_{r}$ representing two instances of the same letter *e* in *#every* are in fact as similar (or as different) as any two other elements. We are just left with token representations, the notion of type of unit remaining out of the representational reach of the system. This makes it difficult to generalize knowledge about phonemes or letters across positions (the so-called dispersion problem: Plaut, McClelland, Seidenberg, and Patterson 1996; Whitney 2001). It is also difficult to align positions across word forms of differing lengths (i.e., the alignment problem: see Davis and Bowers 2004), thus hindering recognition of both shared and different sequences between morphologically-related forms. The failure to provide a principled solution to alignment problems (Daugherty and Seidenberg 1992; Plaut, McClelland, Seidenberg, and Patterson 1996; Seidenberg and McClelland 1989) is particularly critical from the perspective of lexical storage. 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, the alignment of lexical roots in three as diverse pairs of paradigmatically related forms such as English ***walk-walk****ed*, Arabic ***k****a****t****a****b****a****-y****a****k****t****u****b****u*** (‘he wrote’ - ‘he writes’), German ***mach****en-ge****mach****t* (‘make’-‘made’ past participle) requires substantially different processing strategies. Pre-coding any such strategy into lexical representations (e.g. through a fixed templatic structure that separates the lexical root from other morphological markers) would have the effect of slipping in morphological structure directly into the input, thereby 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 exactly what conjunctive coding cannot do.

To our knowledge, there have been three attempts to tackle the issue within a connectionist framework: Recursive Auto-Associative Memories (RAAM; Pollack 1990), Simple Recurrent Networks (SRN; Botvinick and 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 and Plaut’s SRNs and iii) long-term lexical entrenchment for the Sequence Encoder of Sibley and colleagues.

In spite of their considerable differences, all systems share the important feature of modelling 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 particular, Botvinick and Plaut’s paper makes the somewhat paradoxical suggestion that human performance on immediate serial recall develops through direct practice on the task of word repetition. Moreover, short-term memory effects appear to be accounted for in terms of a long-term dynamics dictated by the process of weight adjustment through learning. Although long-term memory effects are known to increase short-term storage capacities, developmental evidence shows that the causal relationship is in fact reversed, with children with higher order short-term memory being able to hold on to new words for longer, thus increasing the likelihood of long-term lexical learning (Baddeley 2007).

We describe here a novel computational architecture for lexical processing and storage. The architecture is based on Kohonen’s Self-Organizing Maps (SOMs; Kohonen 2001) augmented with first-order associative connections that encode probabilistic expectations (so called, Topological Temporal Hebbian SOMs, or T2HSOMs for short; Koutnik 2007; Pirrelli et al. in press; Ferro et al. 2010). T2HSOMs mimic 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. T2HSOMs define an interesting class of general-purpose memory models for serial order, exhibiting a non-trivial interplay between short-term and long-term memory processes. At the same time, they simulate incremental processes of topological self-organization whereby lexical sequences are arranged in maximally predictive hierarchies exhibiting interesting morphological structures.

3. Topological Temporal SOMs

T2HSOMs are grids of topologically organized memory nodes with dedicated sensitivity to time-bound stimuli. Upon presentation of an input stimulus, all map nodes are activated synchronously, but only the most highly activated one, the so-called Best Matching Unit (BMU), wins over the others. Figure 1 illustrates two chains of BMUs triggered by the input German forms *gemacht* and *gelacht* (‘made’ and ‘laughed’ past participle) exposed to a 20x20

nodes map one letter at a time. In the Figure, each node is labelled with the letter the node is most sensitive to after training. Pointed arrows represent temporal connections linking two consecutively activated nodes. The thickness of each arrow gives the strength of the temporal connection. Finally, arrows depict the temporal sequence of node exposure (and node activation), starting from the beginning-of-the-word symbol '#' (anchored in the top left corner of the map) and ending with '$'.

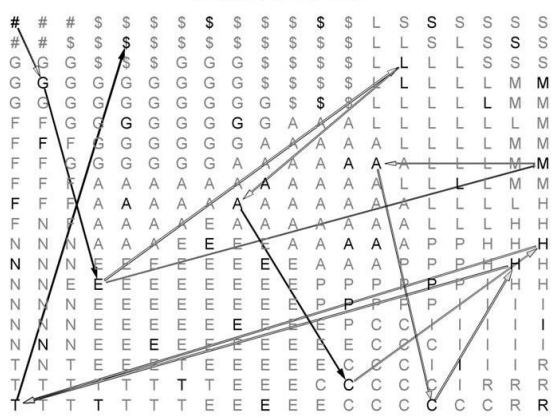

Figure 1 – BMU activation chains for *gemacht-gelacht*

Dedicated sensitivity and topological organization are not wired-in on the map. Neighbouring nodes become increasingly sensitive to letters that are similar in both encoding and distribution through drilling.

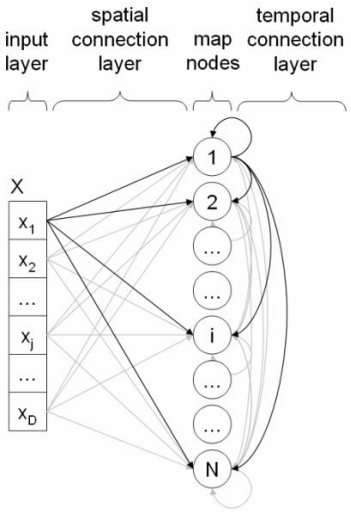

Figure 2 - Outline architecture of a T2HSOM

Figure 2 offers the architecture of a T2HSOM. Each node in the map is connected with all elements of the input layer through communication channels with no time delay, whose strength is modified through training. Connections on the temporal layer, on the other hand, are updated with a fixed one-step time delay, based on activity synchronization of the BMU at time $t-1$ and the BMU at time t. It is important to appreciate at this juncture that, unlike classical conjunctive representations in either Simple Recurrent Networks (Elman 1991) or Recursive SOMs (Voegtlin 2002), where both order and item information is collapsed on the same layer of connectivity, T2HSOMs keep the two sources of information stored on separate (spatial and temporal) layers, which are trained according to independent principles. The aspect has interesting repercussions on issues of order-independent generalizations over symbol types and goes a long way to addressing both dispersion and alignment problems in word matching.

3.1 Memory structures and memory orders

Through repeated exposure to word forms encoded as time series of letters, a T2HSOM shows a tendency to dynamically store strings as trie-like graphs, eliminating prefix redundancy and branching out when two (or more) different nodes are alternative continuations of the same history of past activated nodes (Figure 1). This lexical organization accords well with cohort models of lexical access (Marslen Wilson 1987) and is in keeping with a wide range of empirical evidence on human word processing and storage: i) development of minimally-entropic forward chains of linguistic units, enhancing predictive and anticipatory behaviour in language processing (Altmann and Kamide 1999; Federmeier 2007; Pickering and Garrod 2007); ii) frequency-based competition between inflected forms of the same lexical base (e.g. *brings* and *bringing*) (Hay 2001; Ford, Marslen-Wilson and Davis 2003; Lüdeling and De Jong 2002; Moscoso del Prado Martín, Bertram, Häikiö, Schreuder and Baayen 2004); iii) simultaneous activation of false morphological friends (e.g. *broth* and *brother*) (Frost et al. 1997; Longtin et al. 2003; Rastle et al. 2004; Post, Marslen-Wilson, Randall and Tyler 2008).

It can be shown that trie-like memory structures maximize the map's expectation of upcoming symbols or, equivalently, minimize the entropy over the set of transition probabilities between consecutive BMUs. This is achieved through a profligate use of memory resources, whereby several nodes are recruited to be most sensitive to contextually specific occurrences of the same letter.

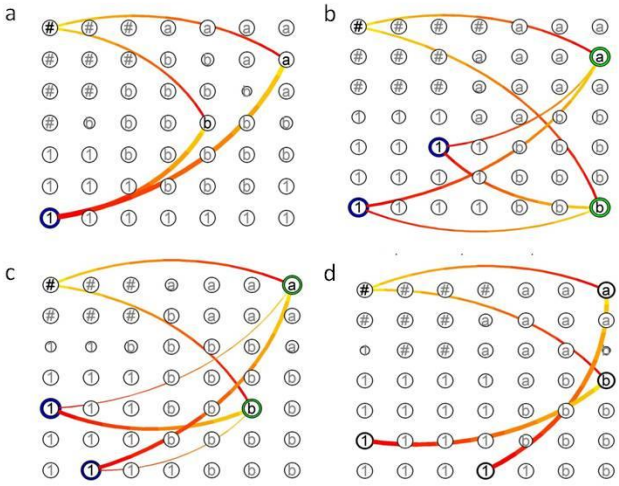

Figure 3 – Stages of chain dedication through learning

Figure 3 illustrates how this process of incremental specialization unfolds through training. For simplicity we are assuming that the map is trained on two strings only: *#a1* and *#b1*. Panel a) represents an early stage of learning, when the map recruits a single BMU for the symbol *1* irrespective of its embedding context. After some more learning epochs, two BMUs are recruited after an *a* or a *b* through equally strong connections (Panel b). Connections get increasingly specialized in Panel c), where the two *1* nodes are preferentially selected by either context. Finally, Panel d) illustrates a stage of dedicated connections, where each *1* node is selected by one specific left context only. This 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 *r* is equal to 0.

4. Emergent Morphological Structure

To what extent do we find morphological structure in a lexical map organized according to the principles sketched above? We observe a straightforward correlation between morphological segmentation and topological organization of BMUs on the map: word forms sharing sub-lexical constituents tend to trigger chains of identical or neighbouring nodes.

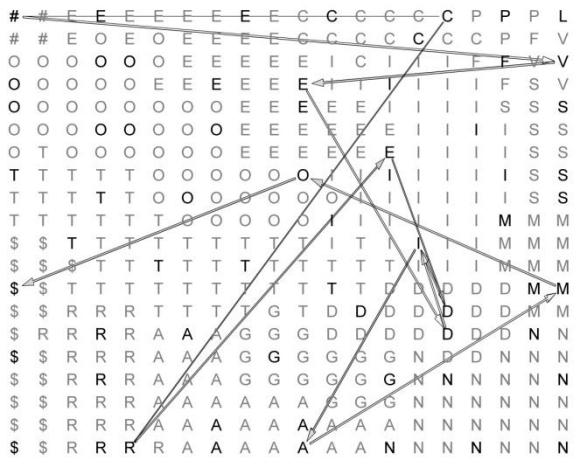

Figure 4 – BMU activation chains for *crediamo-vediamo*

The map distance between BMUs triggered by identical morphemic constituents of two morphologically-related forms is expected to be shorter than the map distance between BMUs activated by morphologically heterogeneous constituents. In a nutshell, topological distance is a function of morphological proximity. In traditional approaches to word segmentation, this is equivalent to aligning morphologically-related word forms by morphological structure. As chains of activated nodes encode time sequences of symbols, T2HSOMs can be said to enforce alignment through synchrony.

To illustrate, we trained three different instances of a T2HSOM on Italian, German and Arabic verb forms. Figure 4 plots the activation chains of the present indicative forms *vediamo* ('we see') and *crediamo* ('we believe') on a 20x20 nodes Italian map, trained on 32 Italian verb forms. The chains are clearly separated on the roots *cred-* and *ved-*, but converge as soon as more letters are shared by the two forms. Eventually the substring *-iamo* activates a unique BMU chain. We take this to mean that the substring is recognized by the map as encoding the same type of inflectional ending. Note that the shared substring *-iamo* takes different positions in the two forms, starting from the forth letter in *vediamo* and from the fifth letter in *crediamo*. In traditional positional coding, this raises an alignment problem. In our map, *-iamo* receives a converging topological representation, as order information is relative and time-dependent rather than absolute.

German past participles provide a case of discontinuous morphological structure. Let us turn back to Figure 1 above. Note that *gemacht* and *gelacht* share the same sequence of BMUs for *ge-*, but they part on the roots *mach-* and *lach-* to eventually meet again upon recognition of the ending *–t*. This is expressed in terms of topological distance between BMUs in Figure 5, giving the per-node topological distance of the BMU chains for *gemacht* and *gelacht*.

	#	G	E	L	A	C	H	T	$
#	0.00	0.24	0.43	0.44	0.40	0.58	0.59	0.53	0.25
G	0.24	0.00	0.37	0.42	0.35	0.52	0.54	0.48	0.24
E	0.43	0.37	0.00	0.47	0.30	0.37	0.48	0.30	0.40
M	0.55	0.52	0.49	0.31	0.38	0.42	0.25	0.59	0.48
A	0.45	0.41	0.39	0.27	0.10	0.38	0.31	0.51	0.38
C	0.63	0.57	0.43	0.49	0.40	0.06	0.34	0.46	0.57
H	0.58	0.53	0.46	0.38	0.37	0.33	0.03	0.54	0.51
T	0.53	0.48	0.30	0.59	0.42	0.40	0.56	0.00	0.52
$	0.25	0.24	0.40	0.36	0.34	0.53	0.51	0.52	0.00

Figure 5 – Topological distance matrix for *gemacht-gelacht*

Besides identical nodes for *ge-* and *–t*, the matrix shows that morphological structure is inherently graded on morpheme boundaries, with the topological distance between the roots narrowing down as the shared suffix gets closer, in keeping with psycholinguistic evidence on word processing (Hay and Baayen 2005).

	#	G	E	S	P	I	E	L	T	$
#	0.00	0.24	0.43	0.46	0.55	0.63	0.57	0.44	0.53	0.25
S	0.51	0.50	0.53	0.06	0.41	0.44	0.55	0.26	0.61	0.44
P	0.52	0.47	0.35	0.44	0.06	0.31	0.29	0.38	0.37	0.46
I	0.63	0.58	0.48	0.47	0.26	0.00	0.38	0.42	0.48	0.56
E	0.57	0.51	0.16	0.55	0.33	0.38	0.00	0.49	0.29	0.52
L	0.44	0.42	0.45	0.24	0.36	0.42	0.49	0.00	0.54	0.37
E	0.42	0.36	0.08	0.43	0.33	0.41	0.15	0.37	0.34	0.37
N	0.48	0.42	0.24	0.57	0.44	0.51	0.34	0.52	0.24	0.46
$	0.39	0.36	0.40	0.28	0.36	0.43	0.46	0.24	0.49	0.14

Figure 6 – Topological distance matrix for *spielen-gespielt*

A case of root-alignment in German lexically-related forms is illustrated in Figure 6, showing the per-node distance between *spielen* and *gespielt*. Once more, this would be out of reach of positional coding.

More difficult cases of root-alignment arise in the context of Semitic morphologies, where the relative position of the letters shared by lexically-related forms vary dramatically, as in *kataba* vs. *yaktubu*, respectively the perfective and imperfective forms of the verb triliteral root *ktb* ('write'). An interesting related question is to what extent the activation chains corresponding to Arabic perfective and imperfective forms are successful in representing the morphological notions of triconsonantal root and interdigitated vowel pattern. The problem is not trivial, as discontinuous morphological patterns are known to be beyond the reach of chaining models for serial order. Given two forms like *kataba* ('he wrote') and *hadama* ('he shattered') for example, vowels in the two strings are all preceded by different left contexts.

	#	K	a	T	a	B	a
#	0.00	0.41	0.28	0.63	0.55	0.46	0.49
H	0.44	0.22	0.37	0.32	0.30	0.26	0.39
a	0.30	0.34	0.05	0.53	0.25	0.40	0.20
D	0.57	0.31	0.51	0.19	0.28	0.30	0.47
a	0.60	0.43	0.34	0.32	0.05	0.44	0.24
M	0.49	0.21	0.47	0.30	0.39	0.17	0.53
a	0.53	0.48	0.25	0.51	0.18	0.53	0.05

Figure 7 – Topological distance matrix for *kataba-hadama*

Figure 7 illustrates the solution offered by a T2HSOM to the problem. The three *a*'s in the perfective vowel pattern are given dedicated representations on the map, triggering differently located BMUs. Not only is the map able to discriminate between three different instances of the same symbol (*a*) in the same string (*kataba*), but it can also align each such *a* with its homologous *a* in another morphologically-related form (*hadama*). In fact, this seems to be a necessary step to take if we want the map to get a notion of the Arabic perfective vowel pattern.

To understand how this is possible, observe that temporal information is not limited to information about the actually occurring left context. The BMU activated by the symbol *a* in the input string *#ha* at time *t* receives support, through temporal connections, from all nodes activated at time *t-1*. The nodes include, among others, the *k* node, which competes with the *h* node at time *t-1* as it receives temporal support from the # node activated at time *t-2* (due to the existence of *#ka* in *kataba*). By reverberating simultaneous activation of competing nodes to an ensuing state, the map can place *a* nodes triggered by *#ka* and *#ha* in the same area, as they share a comparatively large portion of pre-synaptic support. In general, the mechanism allows the map to keep together nodes activated by letters in the same position in the string.

5. Lexical access and recall

So far, we considered chains of BMU activation based on exposure to time-bound sequences of letters. By inspecting activation chains, we can tell whether the map recognizes an input signal as a specific sequence of symbols or not. This is not trivial and requires both sensitivity to letter codes and the capacity of anticipating upcoming symbols on the basis of already seen symbols. Nonetheless, it says little about issues of lexical storage *per se*. How do we know that the map has actually stored the sequence it is able to recognize?

We can model lexical recall as the task of reinstating a sequence of letters from the integrated pattern of activation of a map that has just seen that sequence. Recall that a form is exposed to the map one letter at a time. At each time tick, each letter leaves an activation pattern that accumulates in the map short-term buffer. When the whole form is shown, the map's short-term buffer will thus retain the concurrent activation of all letters forming the just seen word (Figure 8).

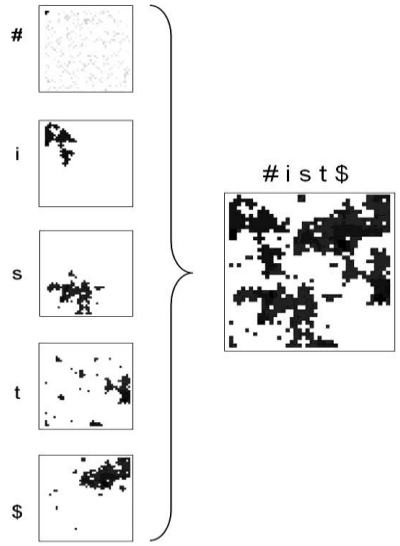

Figure 8 – Per-letter and concurrent activation for *#ist$*

We may eventually feed this pattern back into the map and ask the map to recall from it the expected sequence of letters. Note that this is a considerably more difficult task than activating a specific node upon seeing a particular letter. A whole word integrated pattern of activation is the lexical representation for that word. If the map is able to accurately encode letters and their order of appearance, it will be successful in accessing and retrieving the whole word from its long-term store.

To assess the capacity of a T2HSOM to develop, access and retrieve lexical representations, we trained a 40x40 map on 5000 Italian word forms, sampled from the book *The Adventures of Pinocchio* by Collodi. We then probed the memory content of the map on two test sets: the entire set of "training" word tokens (about 1050 different form types), and a sample of about 250 unseen inflected forms of all verbs that are found in the training set in at least one other form. No frequency information was given for the latter "testing" set.

Results of the experiments are shown in Figure 9 in terms of per-word and per-letter accuracy over types and tokens.

Italian			accuracy	
			% types	% tokens
recognition	training set	per word	99.2	99.7
		per letter	100	100
	testing set	per word	99.6	99.6
		per letter	100	100
recall	training set	per word	97.3	98.8
		per letter	99.1	99.6
	testing set	per word	75.7	75.7
		per letter	95.1	95.1

Figure 9 – Accuracy results on seen and unseen Italian word forms

German			accuracy	
			% types	% tokens
recognition	training set	per word	99.6	98.5
		per letter	99.6	99.9
	testing set	per word	96.7	96.7
		per letter	99.6	99.6
recall	training set	per word	94.2	97.9
		per letter	98.9	99.6
	testing set	per word	80.7	80.7
		per letter	95.8	95.8

Figure 10 – Accuracy results on seen and unseen German word forms

Figure 10 shows the results of a 40x40 T2HSOM trained on 5000 German word tokens (about 1750 different form types), sampled from three fairy tales by brothers Grimm. The testing set included 150 unseen inflected forms of verbs and nouns that are found in the training set in at least one other form, with no frequency information.

All in all, T2HSOMs show a remarkable capacity of activating appropriate BMUs upon recognition of input letters, both on seen words (training set) and unseen words (testing set). Moreover, they can also recall most such words. In fact more than 97% of the Italian forms and more than the 94% of the German forms in the training set are retrieved accurately through activation of BMUs chains. On both the Italian and German training sets, recall errors strongly correlate with low word frequency and word length effects, with most missed word forms showing frequency values close to 1 (Figure 11).

That more than just storage is involved here is shown by the results on the testing set, assessing the ability of the map to "recall" unseen words. More than 75% Italian unseen words and 80% German unseen words are retrieved accurately, meaning that the maps developed memory traces of expected, rather than simply attested, sequences. T2HSOMs can in fact structure familiar information in a very compact (but accurate) way through shared activation paths, thus making provision for connection chains that are never triggered in the course of training. The effect is reminiscent of what we noted in Figure 3 above, where wider neighbourhoods, typical of early stages of learning, favour profligate and more liberal inter-node connections. Only when the map is free to train neighbouring nodes independently, dedicated paths develop. In the current experimental setting, the map is too small to be able to dedicate a different node to each different context-dependent occurrence of a letter.[1] Fewer nodes are recruited to be sensitive to several different context-dependent 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 3.b) and 3.c), where both the *a* and *b* nodes develop more outgoing connections than those strictly required by the training evidence. Most notably, this is the by-product of the way the map stores and structures lexical information.

Italian training set	frequency		length	
	μ	σ	μ	σ
all words	2.8	7.4	7.0	2.5
correctly recalled words (97.3%)	2.8	7.5	7.0	2.5
wrongly recalled words (2.7%)	1.2	0.4	8.6	2.4
German training set				
all words	2.9	6.7	5.9	2.4
correctly recalled words (94.2%)	3.0	6.9	5.7	2.3
wrongly recalled words (5.8%)	1.1	0.3	8.9	2.7

Figure 11 – Mean value and standard deviation of word form frequency and length for Italian and German training sets.

6. Concluding Remarks and Developments

To date, both symbolic and connectionist approaches to the lexicon have laid emphasis on processing aspects of word competence only, whereby morphological productivity is modelled as the task of outputting a – possibly – unknown word form (say an inflected form like *shook*) by taking as input its lexical base (*shake*). Such a "derivational" approach to word competence (Baayen 2007), however, obscures the interplay between storage and computation, adhering to a view of morphological competence as the ability to play a word game.

Symbolic approaches encode word forms using traditional computational devices for storage, allocation and serial order representation such as ordered sets, strings and the like. These devices provide built-in means of serializing order information through chains of pointers which are accessed and manipulated by independently required recursive algorithms. In classical connectionist architectures (Rumelhart and McClelland 1986), on the other hand, the internal representation of word forms in the lexicon is modelled by the pattern of connections between the hidden and the output layer in a multilayered

[1] A 1600 nodes T2HSOM uses up the 2.5% level of connectivity required to store all forms as dedicated BMU chains.

perceptron mapping lexical bases onto inflected forms (e.g. *go* vs. *went*). Serial order is pre-encoded through dedicated nodes, and the resulting lexical organization appears to be contingent upon the requirements of the task of generating novel forms. In principle, different tasks may impose different structures on the lexicon.

In this paper we took a somewhat different approach to the problem. We assumed that word storage plays a fundamental role in both word learning and processing. The way words are structured in our long-term memory (the lexicon) is key to understanding the mechanisms governing word processing and productivity. This perspective offers a few advantages. First, it allows scholars to properly focus on word productivity (the *explanandum*) as the by-product of more basic memory strategies (our *explanans*) that must independently be assumed to account for fundamental aspects of word learning (including but not limited to storage of word forms). Secondly, it opens up new promising avenues of scientific inquiry by tapping the large body of empirical evidence on short-term and long-term memorization strategies for serial order (see Baddley 2007 for a comprehensive recent overview). Furthermore, it gives the opportunity of using sophisticated computational models of language-independent memory processes (Brown Preece and Hulme 2000; Henson 1998; Burgess and Hitch 1996, among others) to shed light on language-specific aspects of word encoding. Finally, it promises to provide a comprehensive picture of the complex dynamics between computation and memory underlying morphological processing.

Put in a nutshell, the processing of unknown words requires mastering rule-governed combinatorial processes. In turn, these processes presuppose knowledge of the sub-word units to be combined. We argue that preliminary identification of the basic inventory of such units depends on memorization of their complex combinations. The way information is stored thus reflects the way such information is dynamically represented, and eventually accessed and retrieved as patterns of concurrent activation of memory areas. According to the view endorsed here, memory processes have the ability not only to hold information but also to structure and manipulate it.

By exploiting the full potential of T2HSOMs, we can simulate processes of dynamic interaction between short-term and long-term memory processes on a classical memory task like Immediate Serial Recall (Henson 1998; Cowan 2001). Moreover, we can investigate aspects of co-organization of concurrent temporal maps, each trained on different modalities of the same input stimuli. This dynamic is key to modelling pervasive aspects of synchronization of multi-modal sequences in both linguistic (e.g. reading) and extra-linguistic (e.g. visuomotor coordination) tasks (Ferro et al. 2011). Finally, we are in a position to explore emergence of islands of reliability (Albright 2002) in the morphological lexicon to account for processes of analogy-driven generalization on the morphological input.

7. References

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

Altmann, G.T.M., and Kamide, Y. (1999), Incremental interpretation at verbs: restricting the domain of subsequent reference, *Cognition*, 73, 247-264

Baayen, H. (2007), Storage and computation in the mental lexicon, in G. Jarema and G. Libben (eds.), *The Mental Lexicon: Core Perspectives*, Amsterdam, Elsevier, 81-104.

Baddeley, A.D. (1986), *Working memory*, New York, Oxford University Press.

Baddeley, A.D. (2006), Working memory: an overview, in S. Pickering (ed.), *Working Memory and Education*, New York, Academic Press, 1-31.

Baddeley, A.D. (2007), *Working memory, thought and action*, Oxford, Oxford University Press.

Baddeley, A.D., and Hitch, G. (1974), Working memory, in G.H. Bower (ed.), *The psychology of learning and motivation: Advances in research and theory*, New York, Academic Press, 8, 47-89.

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

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

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

Cowan, N. (2001), The magical number 4 in short-term memory: A reconsideration of mental storage capacity, *Behavioral and Brain Sciences*, 24, 87-185.

Daugherty, K., and Seidenberg, M.S. (1992), Rules or connections? The past tense revisited, in *Proceedings of the 14th Annual Meeting of the Cognitive Science Society*, Hillsdale, NJ, Erlbaum.

Davis, C.J., and 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.

Di Sciullo, A. M. and Williams, E. (1987). *On the Definition of Word*, Cambridge, MA: MIT Press.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001), DRC: A Dual Route Cascaded model of visual word recognition and reading aloud, *Psychological Review*, 108, 204-256.

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

Federmeier, K.D. (2007), Thinking ahead: the role and roots of prediction in language comprehension, *Psychophysiology*, 44, 491-505.

Ferro, M., Ognibene, D., Pezzulo, G., and Pirrelli, V. (2010), Reading as active sensing: a computational model of gaze planning in word recognition, *Frontiers in Neurorobotics*, DOI: 10.3389/fnbot.2010.00006,

issn: 1662-5218, 4(6), 1-16.

Ferro, M., Chersi, F., Pezzulo, G., and Pirrelli, V. (2011), Time, Language and Action - A Unified Long-Term Memory Model for Sensory-Motor Chains and Word Schemata, in *Intelligent and Cognitive systems*, P. Kunz (ed.), *ERCIM News*, vol. 84 pp. 27-28.

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

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

Harm, M.W., and 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.

Hay, J.B., and Baayen, R.H. (2005), Shifting paradigms: gradient structure in morphology, *Trends in Cognitive Sciences*, 9, 342-348.

Henson, R.N. (1998), Short-term memory for serial order: The start-end model, *Cognitive Psychology*, 36, 73-137.

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

Koutnik, J. (2007), Inductive Modelling of Temporal Sequences by Means of Self-organization, in *Proceeding of International Workshop on Inductive Modelling* (IWIM 2007), Prague, 269-277.

Lüdeling, A., and Jong, N. de (2002), German particle verbs and word formation, in N. Dehé , R. Jackendoff, A. McIntyre and S. Urban, (eds.), *Explorations in Verb-Particle Constructions*, Berlin, Mouton der Gruyter.

Marslen-Wilson, W. (1987), Functional parallelism in spoken word recognition, *Cognition*, 25, 71-102.

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

McClelland, J.L., and 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.

Miller, G.A. (1956), The magical number seven, plus or minus two: Some limits on our capacity for processing information, *Psychological Review*, 63 (2), 81-97.

Moscoso del Prado Fermìn, M., Bertram, R., Häikiö, T., Schreuder, R., and 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., and Zorzi, M. (2007), Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud, *Psychological Review*, 114(2), 273-315.

Pickering, M.J. and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11, 105-110

Pirrelli, V. (2000), *Paradigmi in Morfologia. Un approccio interdisciplinare alla flessione verbale dellitaliano*, Pisa, Istituti Editoriali e Poligrafici Internazionali.

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

Plaut, D.C., McClelland, J.L., Seidenberg, M.S., and 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., and Tyler, L.K. (2008), The processing of English regular inflections: Phonological cues to morphological structure, *Cognition*, 109, 1-17.

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

Rastle, K., Davis and M.H. (2004), The broth in my brothers brothel: Morpho-orthographic segmentation in visual word recognition, *Psychonomic Bulletin and Review*, 11(6), 1090-1098.

Seidenberg, M.S., and McClelland, J.L. (1989), A distributed, developmental model of word recognition and naming, in A. Galaburda (ed.), *From neurons to reading*, MIT Press.

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

Voegtlin, T. (2001) 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.

Appendix - The *T2HSOM* model

A.1 Short-term dynamics: activation and filtering

In recognition mode, the activation level of the map's *i*-th node at time *t* is:

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

where α and β weigh up the respective contribution of the spatial and temporal layers, and

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

is the normalized Euclidean distance between the input vector *x(t)* at time *t* and the spatial weight vector associated with the *i-th* node, and

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

is the weighted temporal pre-activation of the *i*-th node at time *t* prompted by the state of activation of all *N* nodes of the map at time *t-1*. The BMU at time *t* is identified by looking for the maximum activation level

$$y'_{bmu}(t) = \max_i \{y'_i(t)\}$$

eventually normalized to ensure network stability over time:

$$Y(t) = \frac{Y'(t)}{y'_{bmu}(t)}$$

A.2 Long-term dynamics: learning

In *T2HSOM* learning consists in topological and temporal co-organization.

(i) Topological learning

In classical SOMs, this effect is taken into account by a neighbourhood function centered around BMU. Nodes that lie close to BMU on the map are strengthened as a function of BMU's neighbourhood. The distance between BMU and the *i*-th node on the map is calculated through the following Euclidean metrics:

$$d_i(t) = \sqrt{\sum_{c=1}^{n} [i_c - bmu_c(t)]^2}$$

where *n* is 2 when the map is two-dimensional. The topological neighbourhood function of the *i*-th neuron is defined as a Gaussian function with a cut-off threshold:

$$c_{S,i}(t) = \begin{cases} e^{-\frac{d_i^2(t)}{2\sigma_S^2(t_E)}} & if\ d_i(t) \le \nu_S(t_E) \\ 0 & if\ d_i(t) > \nu_S(t_E) \end{cases}$$

where $\sigma_S(t_E)$ is the topological neighbourhood shape coefficient at epoch time t_E, and $\nu_S(t_E)$ is the topological neighbourhood cut-off coefficient at epoch time t_E.

The synaptic weight of the *j*-th topological connection of the *i*-th node at time *t+1* and epoch t_E, is finally modified as follows:

$$\Delta w_{i,j}(t) = \alpha_S(t_E) \cdot c_{S,i}(t) \cdot [x_j(t) - w_{i,j}(t)]$$

$$w_{i,j}(t+1) = w_{i,j}(t) + \Delta w_{i,j}(t)$$

where $\alpha_S(t_E)$ is the topological learning rate at t_E.

(ii) Temporal learning

On the basis of BMU at time *t-1* and BMU at time *t*, three learning steps are taken:

- temporal connections from BMU at time *t-1* (the *j*-th neuron) to the neighbourhood of BMU at time *t* (the *i*-th neurons) are strengthened:

$$m_{i,j}(t+1) = m_{i,j}(t) + \alpha_T(t_E) \cdot c_{T,i}(t) \cdot [1 - m_{i,j}(t) + \beta_T(t_E)]$$

$$c_{T,i}(t) = e^{-\frac{d_i^2(t)}{2\sigma_T^2(t_E)}}$$

- temporal connections from all neurons but BMU at time *t-1* (the *j*-th neurons) to the neighbourhood of BMU at time *t* (the *i*-th neurons) are depressed as well:

$$m_{i,j}(t+1) = m_{i,j}(t) - \alpha_T(t_E) \cdot [1 - c_{T,i}(t)] \cdot [m_{i,j}(t) + \beta_T(t_E)]$$

$$c_{T,i}(t) = e^{-\frac{d_i^2(t)}{2\sigma_T^2(t_E)}}$$

- temporal connections from BMU at time *t-1* (the j-th neuron) to nodes lying outside the neighbourhood of BMU at time *t* (the *i*-th neurons) are depressed as well:

$$m_{i,j}(t+1) = m_{i,j}(t) - \alpha_T(t_E) \cdot c_{T,i}(t) \cdot [m_{i,j}(t) + \beta_T(t_E)]$$

$$c_{T,i}(t) = \begin{cases} e^{-\frac{d_i^2(t)}{2\sigma_T^2(t_E)}} & if\ d_i(t) \le \nu_T(t_E) \\ 0 & if\ d_i(t) > \nu_T(t_E) \end{cases}$$

(iii) Learning decay

As an epoch ends, an exponential decay process applies to each learning parameter so that the generic parameter *p* at t_E is calculated according to the following equation:

$$p(t_E) = p(0) \cdot e^{-\frac{t_E}{\tau_p}}$$

A complete list of the learning parameters is shown below:

- α_S: learning rate of the topological learning process
- σ_S: shape parameter of the neighbourhood Gaussian function for the topological learning process
- ν_S: cut-off distance of the neighbourhood Gaussian function for the topological learning process
- α_T: learning rate of the temporal learning process
- σ_T: shape parameter of the neighbourhood Gaussian function for the temporal learning

process
- ν_T: cut-off distance of the neighbourhood Gaussian function for the temporal learning process
- β_T: offset of the Hebbian rule within the temporal learning process

(iv) Post processing
At a given epoch t_E, the transition matrix is extracted from the temporal connection weights $m_{i,j}(t_E)$, so that $P_{i,j}(t_E)$ is the probability to have a transition from the *i*-th node to the *j*-th node of the network (i.e., the *j*-th node will be the *BMU* at time *t+1*, given the *i*-th node is the *BMU* at time *t*):

$$P_{i,j} = m_{j,i} \cdot \frac{1}{\sum_{h=1}^{N} m_{h,i}}$$

At the same time the labelling procedure is applied. A label L_i (i.e., an input symbol) is assigned to each node, so that the grapheme-base coding of the *c*-th symbol matches the *i*-th node's space vector best:

$$L_i = \arg\min_c \sqrt{\sum_{j=1}^{D} [x_{c,j}(t) - w_{i,j}(t)]^2} \quad i = 1,...,N$$

A.3 Lexical recall

During the lexical recall task, an activation pattern at time *t* does not die out at time *t+1*, but accrues in the map's short-term buffer. When the whole form is shown, the map's short-term buffer thus retains the integrated activation pattern of all letters of the currently input form. Lexical recall is eventually modeled as the task of restoring the input sequence, by priming the map with the '#' symbol first, followed by the integrated activation pattern. More formally, we define the integrated activation pattern $\hat{Y}\{\hat{y}_1,\ldots,\hat{y}_N\}$ of a word of *k* symbols as the result of choosing

$$\hat{y}_i = \max_{t=2,\ldots,k} \{y_i(t)\} \quad i = 1,...,N$$

Lexical recall is thus modeled by the activation function (see Section A.1 above), with

$$y_{S,i}(t) = \begin{cases} \sqrt{D} - \sqrt{\sum_{j=1}^{D} [x_j(t) - w_{i,j}]^2} & t = 1 \\ \hat{y}_i & t = 2,...,k \end{cases}$$

A.4 Parameter configuration

The experiments shown in the present work were performed using the following parameter configuration:
- 40x40 map nodes
- 30 elements in the input vector (orthogonal symbol character coding)
- 100 learning epochs
- learning rates starting from maximum value (i.e. 1.0), exponentially increasing/decaying over epochs (with a time-constant equal to 25 epochs) according to the training error trend
- spatial shape parameter starting from a value so that the Gaussian function has a gain equal to 90% at the maximum cut-off distance, with no decay over epochs
- temporal shape parameter starting from a value so that the Gaussian function has a gain equal to 20% at the maximum cut-off distance, with no decay over epochs
- cut-off distances starting from the maximum distance between two nodes in the map, exponentially increasing/decaying over epochs (with a time-constant equal to 5 epochs) according to the training error trend
- offset of the Hebbian rule within the temporal learning process starting from 0.01), exponentially increasing/decaying over epochs (with a time-constant equal to 25 epochs) according to the training error trend