IEEE Morocco Section **CiSt'14** Colloquium on Information Science and Technology

PROCEEDINGS

2014 Third IEEE International Colloquium in **I**nformation **S**cience and **T**echnology **(CiSt)**

Tetouan, Morocco, October 20 – 22, 2014

ISBN: 978-1-4799-5979-2

IEEE Catalog Number: CFP1467R-ART

CO-EDITORS

CiSt'14 Conference
Mohammed El Mohajir, USMBA, Fez, Morocco
Mohammed Al Achhab, ENSA, Tetouan, Morocco
Mohamed Chahhou, FSDM, USMBA, Fez, Morocco

ANLP CiSt'14 Iinvited Session
Vito Pirrelli, CNR, Pisa, Italy
Arsalane Zarghili, FST, Fez, Morocco

IoT CiSt'14 Invited Session
Arioua Mounir, ENSA, Tetouan, Morocco
Badreddine El Mohajir, FS, Tetouan, Morocco

IMP CiSt'14 Invited Session
Mohamed ElFar, USMBA,Fez, Morocco

Table of Contents

Word Processing for Arabic Language

A reappraisal of morphology induction through adaptive memory self-organisation strategies

Claudia Marzi **Ouafae Nahli** **Marcello Ferro**
Institute for Computational Linguistics, National Research Council – ILC, CNR Pisa, Italy
claudia.marzi@ilc.cnr.it ouafae.nahli@ilc.cnr.it marcello.ferro@ilc.cnr.it

***Abstract*—Modelling the mental lexicon focuses on processing and storage dynamics, since lexical organisation relies on the process of input recoding and adaptive strategies for long-term memory organisation. A fundamental issue in word processing is represented by the emergence of the morphological organisation level in the lexicon, based on paradigmatic relations between fully-stored word forms.**

Morphology induction can be defined as the task of identifying morphological formatives within morphologically complex word forms.

In the computational framework we propose here (TSOMs), based on Self-Organising Maps with Hebbian connections defined over a temporal layer, the identification/perception of surface morphological relations involves the alignment of recoded representations of morphologically-related input words.

Facing a non-concatenative morphology such as the Arabic inflectional system prompts a reappraisal of morphology induction through adaptive organisation strategies, which affect both lexical representations and long-term storage.

We will show how a strongly adaptive self-organisation during training is conducive to emergent relations between stored word forms, and to high accuracy rates in generalising knowledge of stored words to unknown forms.

***Keywords*—Non-concatenative morphological structure; word recoding and processing; lexical storage and access; topological alignment; synchronisation; self-organising maps.**

I. Introduction

The task of inducing morphological knowledge from lexical data can be defined as the task of singling out morphological formatives from surface word forms. Operationally, the task consists of the following steps: (i) finding structure in word forms, and (ii) grouping word forms on the basis of shared structure. Originally defined by Harris [1] as a battery of "discovery procedures" of unclassified training data on the basis of purely formal algorithms, morphology induction mirrors the interplay between structured representation and the recoding process.

Depending on different algorithms, supervised and unsupervised machine learning models provide different setting for word representation and storage. Supervised algorithms tend to rely on specific assumptions on word representations. Indeed, for most European languages, we can construe a fixed-length vector representation that aligns input words to the right, since inflection in those languages typically involves suffixation and sensitivity to morpheme boundaries. However, this type of representation presupposes considerable knowledge of the morphology of the target language and does not possibly work with prefixation, circumfixation and non-concatenative morphological processes in general.

Conversely, most current unsupervised algorithms model morphology learning as a segmentation task [2], assuming a hard-wired linear correspondence between sub-lexical strings and morphological structure. Once more, non-concatenative morphologies can hardly be segmented into linearly concatenated morphemes.

In line with recent psycholinguistic evidence on peripheral levels of automatic morphology segmentations [3, 4, 5], all approaches to morphology induction must address a fundamental issue: adaptivity to morphological structure, which, in some cases, may require differential sensitivity to parallel, coexisting systems [5]. In this perspective, algorithms for morphology learning should be more valued for their general capacity to adapt themselves to the morphological structure of a target language, rather than for the strength of their inductive morphological bias.

Ideally, the same morphology induction algorithm, with an identical setting of initial parameters and a comparable set of assumptions concerning input representations, should be able to successfully deal with as diverse inflectional systems as, for example, Italian, German and Arabic. We suggest that a principled approach to these issues should be able to replicate some fundamental abilities lying at the heart of the human language faculty: (i) recode and maintain time series of symbolic units (e.g. letters, syllables, morphemes, or words) in the so-called working memory, (ii) transfer and organise these representations in the long-term memory, (iii) map input representations onto lexical representations for access/recall.

II. Recoding and Memory

A fundamental characteristic of the human language faculty is the ability to retain sequences of symbolic items, to access them in recognition and production, and to find similarities and differences among them. Traditionally, lexical acquisition and processing have been modelled in terms of basic mechanisms of human memory for serial order, as proposed in the vast literature on immediate serial recall and visual word recognition (e. g. [6, 7] for detailed reviews).

There is a general problem that all such models have to address and that appears to be crucial for morphology induction: the word alignment issue. The problem arises whenever familiar patterns are presented in novel arrangements, as when speakers of English are able to recognise the word *book* in *handbook*, or Arabic speakers can track down the verb root *k-t-b* in *kataba* ('he wrote') and *yaktubu* ('he writes'). No position-specific letter coding scheme can account for such an ability.

In Davis' [7] spatial encoding, a letter in a string is represented as a two-dimensional signal. The identity of the letter is described as a Gaussian activity function whose maximum value is centred on the letter's actual position and decreases continuously as we move away from that position either rightwards or leftwards. The function defines a confidence level on the position of the letter in question. String matching is continuously weighted by levels of positional confidence, thus enforcing a form of fuzzy matching. However, the approach, as most other psycho-cognitively inspired models such as the "open-bigram coding" model [8], the "start-end" model [9] and the "primacy model" [10] among others, is chiefly recognition-oriented and is not readily amenable to purposes of morphology induction and generalisation.

An altogether different approach, more geared towards lexical acquisition, has recently been proposed in the framework of Temporal Self-Organising Maps (TSOMs), a variant of Self-Organizing Memories (SOMs [11]) augmented with re-entrant Hebbian connections defined over a temporal layer, which can encode probabilistic expectations upon incoming stimuli [12, 13, 14, 15, 16, 17]. TSOMs are grids of artificial memory nodes, which are not wired-in to maximally respond to specific symbols (as customary in the more traditional "conjunctive coding" of multi-layered perceptrons [18]), but can be trained to exhibit dedicated sensitivity to time-bound symbols. Upon presentation of a symbol, all nodes in the map fire concurrently, but only the most highly activated one (Best Matching Unit, BMU) wins out. The approach provides a general framework where word processing and lexical acquisition are implemented as recoding and storage strategies for time-series of symbols, dependent on language-specific factors and extra-linguistic cognitive functions such as lexical organisation, lexical access and recall, input-output representations, and adaptive memory self-organisation (for a detailed description of the model see [15, 17]).

During training, each node develops a dedicated sensitivity to a possibly position-specific letter by incrementally adapting its synaptic weights to recurrent patterns of morphological structure. In fact, 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. Each input word form is represented by a unique time-series of symbols (be they phonological representations or transcription letters), which are administered to a TSOM by encoding one symbol at a time on the input layer.

Upon presentation of each symbol on the input layer, all nodes are activated simultaneously through their input/spatial connections, as a function of their sensitivity to the symbol and familiarity with its embedding context. This implies that an entire pool of nodes, during training, tend to specialise to respond to any specific input symbol, each node in the pool showing higher activity levels than all others when the symbols appear in a particular context. The behaviour is reminiscent of the graded activation function in Davis' spatial coding, but is in fact more directly related to the functional co-activation of pools of neurons selectively responding to the same stimulus type. Co-activation of the same BMUs by different input words reflects the extent to which the map perceives surface morphological relations between fully-stored words. We contend that perception of morphological structure has to do with information sharing and high coactivation levels between word forms.

Fig. 1 provides topological distances (top panel) and co-activation distance (bottom panel) of both *kataba* BMUs and *yaktubu* BMUs responding to the input *yaktubu* and *kataba* respectively, in a TSOM trained on Arabic inflected word forms. The BMU of *k* in *yakbutu* is maximally co-activated when the *k* in *kataba* is shown to the map. This is true also for *t* and *b* BMUs in *yaktubu*, and for the corresponding BMUs of *k-t-b* in *kataba* when *yaktubu* is input. Overall, responses of the two pools of nodes are maximally synchronised when the root *k-t-b* is presented. We take these levels of co-activity response to mean that *k-t-b* are perceived as instantiations of the same consonantal skeleton in both word forms. On the contrary, the two forms are topologically misaligned (Fig. 1, top).

	#	y	a	k	t	u	b	u	$
#	**0.00**	0.69	0.56	0.49	0.68	0.65	0.71	0.71	0.25
k	0.44	0.42	0.32	**0.09**	0.62	0.25	0.32	0.27	0.20
a	0.51	0.18	**0.05**	0.23	0.32	0.51	0.29	0.47	0.35
t	0.71	**0.12**	0.18	0.42	0.22	0.67	0.35	0.61	0.57
a	0.87	0.25	0.35	0.41	0.57	0.53	0.18	**0.42**	0.66
b	0.77	0.24	0.29	0.28	0.57	0.40	**0.05**	0.30	0.54
a	0.59	0.13	0.09	0.33	0.23	0.61	0.33	0.55	**0.45**
$	0.48	0.32	0.24	0.42	**0.20**	0.70	0.49	0.67	0.42

	#	y	a	k	t	u	b	u	$
#	**0.00**	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
k	0.33	0.18	0.26	**0.08**	0.26	0.26	0.26	0.26	0.26
a	0.33	0.26	**0.04**	0.24	0.20	0.24	0.24	0.24	0.24
t	0.33	0.26	0.24	0.18	**0.06**	0.23	0.23	0.23	0.23
a	0.33	0.26	0.06	0.23	0.22	**0.19**	0.23	0.23	0.23
b	0.33	0.26	0.24	0.19	0.23	0.23	**0.05**	0.23	0.23
a	0.33	0.26	0.06	0.23	0.22	0.23	0.23	**0.19**	0.23
$	0.33	0.26	0.24	0.19	0.23	0.23	0.23	0.23	**0.05**

Fig. 1. Best Matching Unit (BMU) distances for the input forms *kataba* and *yaktubu*. # and *$* stand respectively for the "start-of-word" and "end-of-word" symbols. Distance equal to 0 means that exactly the same node is activated. Top panel: Topological distances for the input forms *kataba* and *yaktubu*. Bottom panel: Co-activation level distances for the input forms *kataba* and *yaktubu*.

Since BMUs become sensitive to both nature and timing of an input symbol through training, the Arabic map develops two distinct *k* nodes: one for *k* in first position and one for *k* in third position. When either *k* is shown as an input stimulus, both *k*-nodes fire concurrently but the most contextually specialised one shows stronger activity. This explains co-activity values slightly above zero on the consonantal skeleton. Most notably, node specialisation is the result of language specific patterns repeatedly recurring in input. Positional specialisation of consonantal nodes thus reflects the specific arrangement of consonants in Arabic morphology. More combinatorial morphotactic systems would hardly prompt the same type of sensitivity.

III. Paradigmatic Relations

One of the most prominent issues in word processing is represented by the emergence of the morphological organisation level in the lexicon. In the perspective of adaptive strategies for lexical acquisition and processing based on emergent morphological relations between fully-stored word forms (defined as an *abstractive* approach after Blevins, [19]), paradigmatic relations can be accounted for as the result of long-term entrenchment of neural circuits (chains of time-stamped memory nodes) that are repeatedly being activated by a memory map in the process of recoding input words into a two-dimensional lexical layer.

In previous work [20], we analysed the paradigmatic organisation of the inflectional morphology of Italian and German, by focussing on how different types of related intra- and inter-paradigmatic families induce a strongly paradigm-related co-organisation and co-activation so as to facilitate paradigmatic extension and generalisation. We have defined [16] generalisation as based on recurrent sub-lexical chunks in the lexicon, which can be combined for the recognition and access of novel words. In a TSOM, this requires that deeply entrenched chains of nodes are concurrently activated by morphologically related word forms. In particular, we highlighted how, from a lexical standpoint, TSOMs exhibit a straightforward correlation between morphological segmentation and topological organisation of BMUs. Word forms sharing sub-lexical constituents tend to trigger chains of identical or neighbouring nodes. In other words, we found that – for concatenative morphologies - topological distance (proximity) on the map correlates with morphological similarity. In traditional morpheme-based approaches (see [21], [22], for recent theoretical revisitations) to word segmentation, this is equivalent to topologically aligning morphologically-related word forms by morphemic structure.

Discontinuous morphological formatives - e.g. roots in the Arabic inflectional system – represent a challenge to the notion that identical structures are responded to by topologically adjacent nodes. The root *k-t-b* is, for example, dramatically misaligned in *kataba* and *yaktubu*, and this may keep the nodes responding to the root in two – or more – words far apart on the map.

Interestingly for our present purposes, however, nodes can fire synchronously – with congruent levels of co-activation - even if they are not topologically adjacent (as shown in Fig. 1, bottom panel vs. top panel). Since chains of activated nodes encode time sequences of symbols, TSOMs can be said to enforce alignment through temporal synchrony (co-activation) as well. In fact, co-activation represents the most basic correlate to the notion of similarity in perception, as witnessed by the huge literature on morphological priming ([23, 24] among others). From this perspective, topological proximity is only a by-product of training a map on concatenative morphological structures. As an example, we report in Fig. 2 topological (top panel) and co-activation level (bottom panel) distances for the input forms *macht* and *gemacht*, to highlight how an almost linear morphology, illustrated by the German verb system, is conducive to the development of both topologically-close and strongly co-activated memory chains.

	#	M	A	C	H	T	$
#	**0.00**	0.44	0.69	0.08	0.74	0.72	0.32
G	0.16	0.34	0.65	0.11	0.60	0.55	0.27
E	0.72	0.28	0.34	0.64	0.12	0.45	0.42
M	0.49	**0.06**	0.34	0.41	0.27	0.45	0.20
A	0.67	0.33	**0.03**	0.61	0.46	0.76	0.36
C	0.08	0.36	0.63	**0.00**	0.66	0.66	0.25
H	0.74	0.33	0.45	0.66	**0.00**	0.35	0.46
T	0.72	0.49	0.76	0.66	0.35	**0.00**	0.59
$	0.32	0.14	0.38	0.25	0.46	0.59	**0.00**

	#	M	A	C	H	T	$
#	**0.00**	0.34	0.34	0.34	0.34	0.34	0.34
G	0.34	0.18	0.27	0.27	0.27	0.27	0.27
E	0.34	0.27	0.23	0.25	0.25	0.25	0.25
M	0.34	**0.09**	0.25	0.24	0.25	0.25	0.25
A	0.34	0.27	**0.02**	0.25	0.25	0.25	0.25
C	0.34	0.27	0.25	**0.00**	0.25	0.24	0.24
H	0.34	0.27	0.25	0.25	**0.00**	0.25	0.25
T	0.34	0.27	0.25	0.25	0.25	**0.00**	0.25
$	0.34	0.27	0.25	0.25	0.25	0.25	**0.00**

Fig. 2. Topological (top panel) and co-activation level (bottom panel) distances for the German input forms *macht* and *gemacht*.

On the contrary, for those (non-concatenative) morphologies where structures are misaligned, perception of internal structure is better correlated with high levels of co-activation than with topological proximity of BMU nodes. To investigate these issues, we ran a few experiments on a portion of the Arabic inflectional lexicon, and compared our results with similar experiments on the German verb system. By controlling a number of quantitative parameters, we could assess, on a fair comparative basis, a range of subtle differences between the ways the two morphological systems are processed and organised. Some of these results are reported in the following section.

IV. Materials, Method and Results

First, we selected fifty of the most frequent verb paradigms (from *A frequency Dictionary of Arabic* [25]). The set contains verbs from various inflectional classes including sound-regular, geminated (i.e. the second and the third consonant root-consonants are similar), or "hamzated" (containing a *hamza* as any of the three root consonants). A few selected verbs are weak-assimilated (i.e. the first verbal consonants is *wāw*), weak-hollow (i.e. the second verbal consonant is *wāw* or *yā'*) or weak-defective (i.e. the third verbal consonant is *wāw* or *yā'*). For each paradigm – whenever attested – we selected up to 12 distinct inflected forms, namely the first, second and third masculine singulars and plurals, for both the perfective and imperfective. Finally, for each verb, we added 3 more derivationally-related forms: the *maṣḍar* (i.e. the verbal noun), the agent noun, the object noun. This made our final set of Arabic forms comparable with German and Italian data used in previous experiments, where we included, besides present indicative and past tense forms, the infinitive, past participle and present participle.

All sampled forms (687) were fully vocalised and orthographically transcribed according to a normalised version of Buckwalter's transliteration system [26]. They were encoded as strings of lower-case and upper-case alphabetic and non-alphabetic ASCII characters (e.g. '?' for *hamza*, '$$' for the *sh* sound as in the English pronoun *she*) starting with '#' (i.e. the start-of-word symbol) and ending with '$' (i.e. the end-of-word symbol). A few special Arabic character-diacritic combinations (e.g. lengthened vowels) were encoded as digraphs of lower-case and upper-case letters (e.g. '#,k,aA,n,a,$' for كَانَ), processed as one symbol by the map. All symbols were encoded on the map's input layer as mutually orthogonal binary vectors.

In Experiment 1, we trained two parametrically different instances of a 40 x 40 nodes map, for five repetitions each, on the 687 word types. The two parametric settings define different degrees of sensitivity of the map to the timing of each symbol, i.e. to its ordered position in the string. For convenience, we will refer to the more time-sensitive maps as "temporally-biased" (TB), and to the less time-sensitive maps as "balanced" (B). To simulate low-level memory processes for serial order and their impact on morphological organisation, only information about raw forms was provided during training. Each input word was administered to a TSOM one symbol at a time, with memory of past symbols being reset upon presentation of '#'. At each training epoch, input forms were presented to the map in random order 5 times each, for a total number of 3,435 presentations per epoch. A map's full training consisted of 100 epochs.

After training, we tested the memory content of the maps and probed the internal organisation on two tasks: word RECODING and word RECALL. The task of RECODING consists in testing the accuracy of the map's activation on input forms. An input word is recognised correctly if each BMU in the activation chain is correctly associated with the current input symbol. Errors are counted when an input symbol activates a BMU associated with a different symbol. Unlike word recognition, which mostly depends on the current input stimulus, word RECALL entirely depends on the memory accuracy of a map, and simulates the reverse process of retrieving a sequence of symbols from its chain of BMUs. This is achieved through spreading of activation from the start-of-word node ('#') through the nodes making up the chain. At each time step, the map outputs the individual symbol associated with the currently most highly-activated node. The step is repeated until the node associated with the end-of-word symbol ('$') is output. Errors occur when the map misrecalls one or more symbols in the input string, by either replacing it with a different symbol or by outputting correct symbols in the wrong order. Partial recall, i.e. the correct recall of only a substring of the target word (e.g. '#,k,a,t,a,b,$' for '#,k,a,t,a,b,a,$'), is also counted as an error. Results on both tasks are provided in Table 1.

In Experiment 2, 45 word forms were left out of the original training set of Experiment 1 and kept for testing under the constraint that each test form must contain a stem which is attested in at least one other form of the training set. According to the same protocol of Experiment 1, five different repetitions of both TB and B maps were then trained on the outstanding 642 forms. Once more, trained maps were evaluated on the tasks of word recoding and word recall for the 45 test forms. For each parametrical setting, results (averaged over five repetitions) for the two tasks of recoding and recall are provided in Table 1.

TABLE I

Experimental results for balanced (B) and time-biased (TB) maps	Scores			
	Recoding	*s. dev*	*Recall*	*s. dev*
Experiment 1: B map	100%	0	98.98%	0.97
Experiment 1: TB map	100%	0	99.24%	0.52
Experiment 2: B map (training)	100%	0	99.22%	0.60
Experiment 2: B map (test)	100%	0	89.78%	3.30
Experiment 2: TB map (training)	100%	0	99.24%	0.55
Experiment 2: TB map (test)	100%	0	89.78%	1.20

Accuracy scores are fairly high for all the tasks and for the two different parametric configurations; it should be appreciated that for the slightly temporally-biased maps (TB) recall results on novel words are more stable across map instances, meaning that a more systematic use of temporal connections strengthens the map's propensity to accept novel words. As pointed out in the previous section, intra- and inter-paradigmatic relations occur to facilitate paradigmatic extension and generalisation. Specifically, our evidence emphasises that the underlying structure of Arabic verb forms requires sensitivity to both time invariant symbol encoding of the root consonant skeleton (intra-paradigmatic relations) and a position-sensitivity to time-bound instances of the same vowel symbol (inter-paradigmatic relations). A temporal bias (TB map) gives an additional support to the inter-paradigmatic relations, allowing different instances of the same vowel to be distinguished, as illustrated by the example reported in Fig. 3.

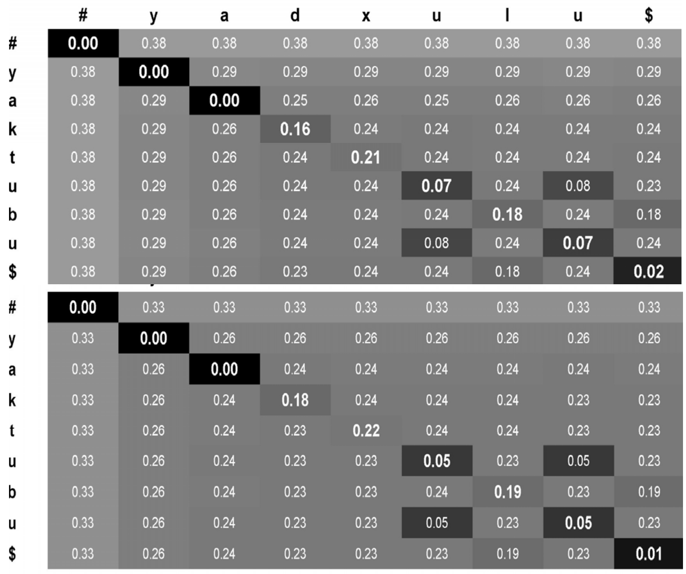

Fig. 3. Co-activation level distances in a TB map (top panel) and a B map (bottom panel) for the input forms *yadxulu* and *yaktubu*.

To verify the overall internal organisation of Arabic TSOMs after training, we assessed how well trained maps perceive the amount of morphological structure shared by all inflected forms of each verb (intra-paradigmatic relations). Averaged results for all 50 paradigms (ordered by graded distance values) are plotted in Fig. 4 for the two map types (white dots for B maps and dark dots for TB maps). On average, less regular paradigms are perceived as internally less coherent (less co-activated) than more regular ones. We observe a graded perception of morphological structure, which emerges from intra-paradigmatic relevant formal redundancy.

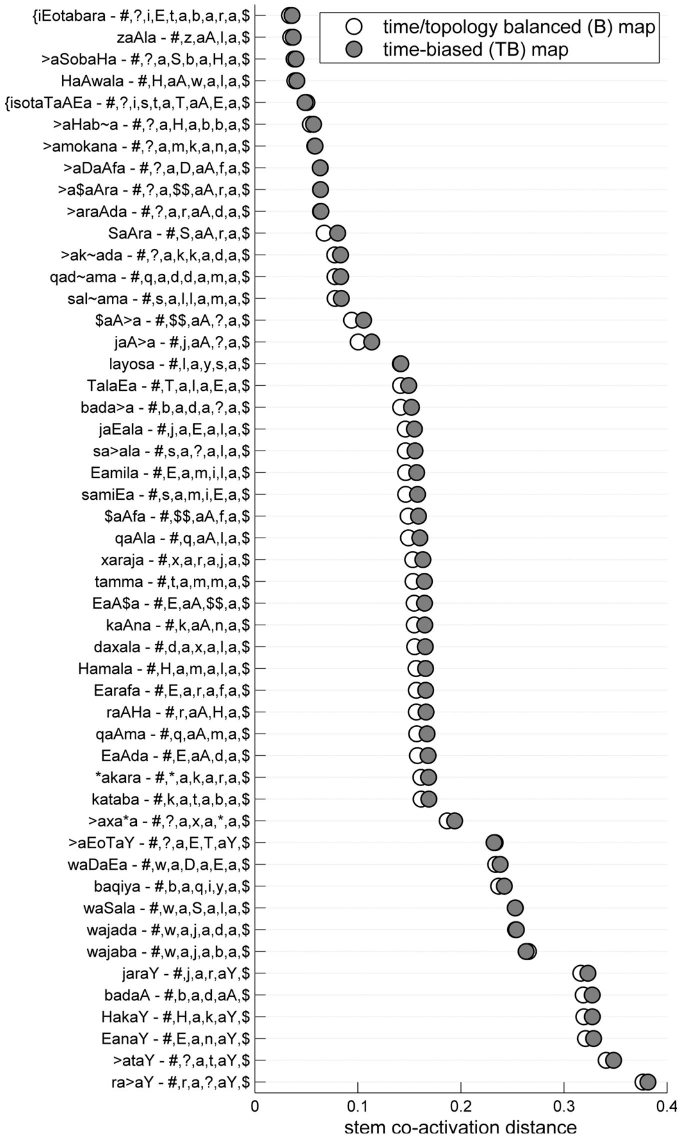

Fig. 4. Internal paradigmatic co-activation for B maps (white dots) and TB maps (dark dots)

The result has two interesting implications. First, TSOMs demonstrably develop a notion of "verb stem", as a distributed pattern of nodes showing persistent co-activation levels across all forms of the same verb. An example of such a shared distributed activation pattern is shown in Fig. 5 for the verb ra>aY (رَأى).

Secondly, this notion is the dynamic result of recoding through training, as shown by the graph in Fig. 6, plotting the amount of intra-paradigmatic coherence that a map is able to perceive over training epochs. As training progresses, TSOMs are more and more able to find out what is common and what is different within each set of paradigmatically-related verb forms.

Finally, distributed activation patterns such as the one in Fig. 5 prove to be able to keep track of elements of the consonantal skeleton which are not attested in all forms of the same paradigm (e.g. *wāw* in weak-hollow paradigms), by assigning them lower levels of shared activation.

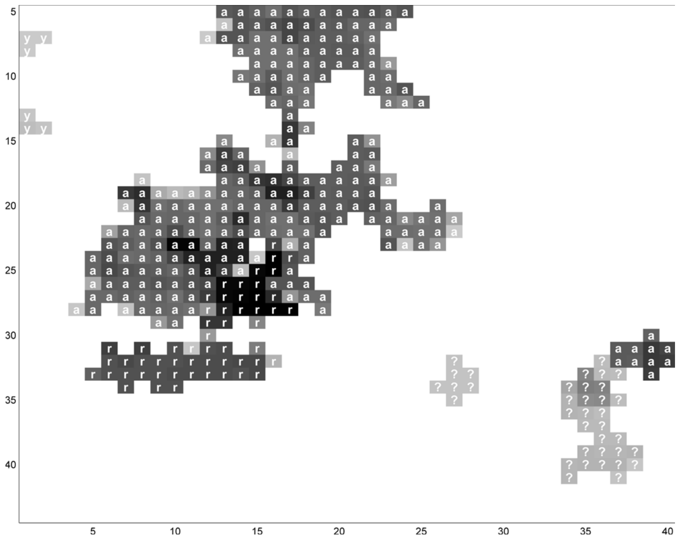

Fig. 5. Shared distributed activation patterns for all trained forms of *ra>aY.*

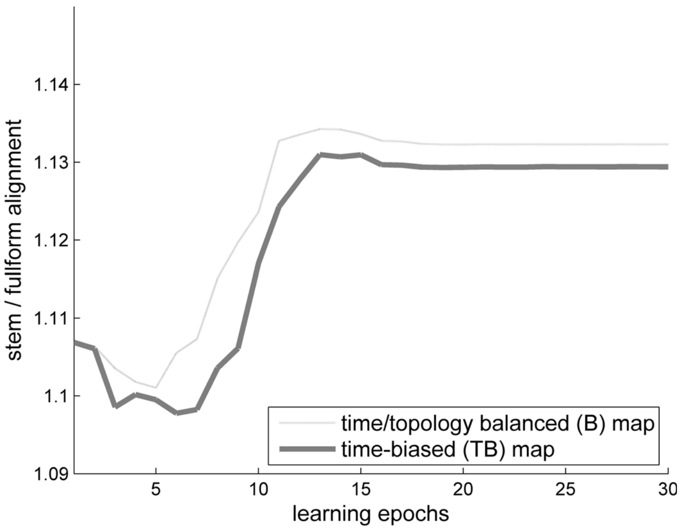

Fig. 6. Intra-paradigmatic coherence over training epochs, for B maps (thin line) and TB maps (thick line).

V. General Discussion and Concluding Remarks

TSOMs like other computational models of serial memory are known to exhibit a learning bias towards selective specialisation of first-order Markovian chains of memory nodes, arranged in tree-like structures.

Arabic inflectional morphology represents a challenge to this chaining bias, due to the chiefly non-concatenative nature of consonantal roots and vowel patterns, and the concurrent presence of prefixes and suffixes. The present work offers the opportunity to comparatively assess the performance of TSOMs on different morphological systems, under controlled parametric and input conditions.

TSOMs prove to be extremely effective in learning a real portion of the Arabic verb system (including a few deverbal nouns), achieving accuracy levels which are remarkably close to those obtained on concatenative languages such as Italian and German (for the latter, see table II). This result was obtained by running experiments with comparable amounts of training data and memory resources (1600 memory nodes exposed to about 700 word types). It should be emphasised that the rate of memory compression achieved by Arabic TSOMs is slightly higher than the compression rate of German (and Italian) maps, measured as the ratio between the number of nodes required to represent the set of input words in a word tree and the number of BMUs needed for a TSOM to process the same set of words.

TABLE II

Experimental results for balanced (B) and time-biased (TB) maps	Scores			
	Recoding	*s. dev*	*Recall*	*s. dev*
Experiment 3: B map (training)	100%	0	99.60%	0
Experiment 3: B map (test)	100%	0	100%	0
Experiment 3: TB map (training)	100%	0	99.38%	0.31
Experiment 3: TB map (test)	100%	0	95.20%	1.10

Our main goal here was to assess to what extent a TSOM exposed to Arabic input is able to effectively store and recall verb forms by acquiring principles of their morphological organisation. In previous work, we showed that morphological organisation is the by-product of the topological arrangement of memory nodes on the map. Chains of nodes responding to the same stem or affix are either overlapping or are located at a close distance on the map. By measuring the distance between nodes responding to the same symbol input, we could assess the level of perception of shared morphological structure by a trained map.

As a general point, Arabic morphology prompts a different and somewhat unexpected type of organisation. The highly non-linear nature of Arabic morphology is not conducive to the development of topologically-close chains. Nonetheless, effective organisation of memory nodes is achieved by different means: namely, by their propensity to respond to the same symbol at different positions in time (CO-ACTIVATION). Generally, a node that is selectively sensitive to a particular symbol in a specific ordered position reaches high level of co-activity when the same symbol is shifted by few positions. This is not always true, but it depends on the type of sequences which are shown to the map during training. In the end, repeated exposure to the underlying structure of Arabic word forms enforces both a mild sensitivity to time invariant symbol encoding (intra-paradigmatic relations) and a position-sensitivity to time-bound instances of the same vowel symbol (inter-paradigmatic relations). In fact, it should be appreciated that the map retains information about the relative order of symbols in a string, as witnessed by the high accuracy rates in the task of word recall, which requires fine-grained sensitivity to time. Furthermore, high accuracy rates in generalising knowledge of stored words to unknown forms are possible only if the map develops time-bound expectations over up-coming symbols. To explore this issue further, in our experiments we modulated a map's sensitivity to time-invariance through different parametrical settings.

Notably, recall of unknown forms has to do with paradigm induction, and requires considerable flexibility in perceiving/co-activating the word to be recalled on the basis of other morphologically-related, stored word forms.

To sum up, perception of nonlinear morphological structures requires more complex processing and storage strategies than simple chaining or positional ordering. In TSOMs, distributed populations of memory nodes can be trained to selectively respond to either time-invariant or context-sensitive recoding of symbols. This appears to provide a more flexible and effective computational approach to Arabic word processing than more traditional approaches.

Computer simulations of lexical processing and storage provide a methodological middle ground for testing models of word acquisition. Through careful data analysis of the computational behaviour of TSOMs, we gained specific insights into issues of paradigmatic acquisition and morphological relations between fully-inflected word forms. Since words are treated like input stimuli producing a change in the activation state of the map, processing and memorising words are modelled as two sides of one coin.

REFERENCES

[1] Z. S. Harris, "From phoneme to morpheme", Language, 31(2), pp. 190-222, 1955.

[2] H. Hammarström, and L. Borin, "Unsupervised learning of morphology", Computational Linguistics, 37(2), pp. 309-350, 2011.

[3] D. Crepaldi, K. Rastle, M. Coltheart, and L. Nickels, "'Fell' primes 'fall', but does 'bell' prime 'ball'? Masked priming with irregularly-inflected primes", Journal of Memory and Language, 63, pp. 83-99, 2010.

[4] K. Rastle, and M. H. Davis, "Morphological decomposition based on the analysis of orthography", Language and Cognitive Processes, 23, pp. 942-971, 2008.

[5] H. Velan, and R. Frost, "Words with and without internal structure: What determines the nature of orthographic and morphological processing?" Cognition, 118, pp. 141-156, 2011.

[6] R. N.A. Henson, "Coding position in short-term memory", International Journal of Psychology, 34 (5-6), pp. 403-409, 1999.

[7] C. J. Davis, "The spatial coding model of visual word identification", Psychological Review, 117.3, pp. 713-758, 2010.

[8] J. Grainger, W. van Heuven, "Modeling letter position coding in printed word perception", The mental lexicon, 1-24. New York, Nova Science 2003.

[9] R. N.A. Henson, "Short-term memory for serial order: the start-end model", Cognitive Psychology, 36(2), pp. 73-137, 1998.

[10] M. P.A. Page, D. Norris, "The primacy model: a new model of immediate serial recall", Psychological Review, 105, pp. 761-781, 1998.

[11] T. Kohonen, "Self-organizing maps", Berlin-Heidelberg: Springer-Verlag, 2001.

[12] J. Koutnik, “Inductive Modelling of Temporal Sequences by Means of Self-organization” Proceeding of International Workshop on Inductive Modelling pp. 269-277, 2007 [IWIM 2007, Prague].

[13] M. Ferro, G. Pezzulo, and V. Pirrelli, “Morphology, Memory and the Mental Lexicon”, in V. Pirrelli (ed.), Lingue e Linguaggio, vol. IX(2), Interdisciplinary aspects to understanding word processing and storage, Bologna: Il Mulino, pp. 199-238, 2010.

[14] V. Pirrelli, M. Ferro, and B. Calderone, “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, pp. 135-157, 2011.

[15] M. Ferro, C. Marzi, and V. Pirrelli, “A Self-Organizing Model of Word Storage and Processing: Implications for Morphology Learning”, Lingue e Linguaggio 2, pp. 209-226, 2011.

[16] C. Marzi, M. Ferro, and V. Pirrelli, “Prediction and Generalisation in Word Processing and Storage”, 8th Mediterranean Morphology Meeting Proceedings on Morphology and the architecture of the grammar, pp. 113-130, 2012 [MMM8, Cagliari].

[17] C. Marzi, M. Ferro, V. Pirrelli, “Word alignment and paradigm induction” Lingue e Linguaggio XI, 2, pp. 251-274, 2012.

[18] D. Rumelhart, and J. McClelland, “On learning the past tense of English verbs”, in D.E. Rumelhart and J. McClelland (eds.), Parallel distributed processing: Explanations in the microstructure of cognition, The MIT Press, pp. 216-271, 1986.

[19] J. P. Blevins, “Word-based morphology”, Journal of Linguistics, 42, pp. 531-573, 2006.

[20] C. Marzi, M. Ferro, C. Caudai, and V. Pirrelli. “Evaluating Hebbian Self-Organizing Memories for Lexical Representation and Access”, Proceedings of 8th International Conference on Language Resources and Evaluation, pp. 886-893, 2012 [ELRA - LREC 2012, Malta].

[21] M. Halle, and A. Marantz, “Distributed Morphology and the pieces of inflection”, in K. Hale and S. J. Keyser (eds.), The view from building 20, Cambridge (MA): MIT Press, pp. 111-176, 1993.

[22] D. Embick, and M. Halle, “On the Status of stems in morphological theory”, in T. Geerts, I. van Ginneken, and H. Jacobs (eds.), Romance Languages and Linguistics Theory 2003, Amsterdam: John Benjamins, pp. 37-62, 2005.

[23] M. S. Seidenberg, G. S. Waters, M. Sanders, and P. Langer, “Pre- and postlexical loci of contextual effects on word recognition”, Memory and Cognition, 12(4), pp. 315-328, 1984.

[24] K. I. Forster, “The pros and cons of masked priming”, Journal of Psycholinguistic Research, 27(2), pp. 203-233, 1998.

[25] T. Buckwalter, and D. Parkinson, “A Frequency Dictionary of Arabic”, New York: Routledge, pp. 566-567, 2011.

[26] http://www.qamus.org/transliteration.htm