7

Learning Paradigms in Time and Space: Computational Evidence from Romance Languages*

VITO PIRRELLI
(Institute of Computational Linguistics ‘A. Zampolli’ (ILC),
National Research Council (CNR), Pisa),

MARCELLO FERRO
(Institute of Computational Linguistics ‘A. Zampolli’ (ILC),
National Research Council (CNR), Pisa), and

BASILIO CALDERONE
(MoDyCO – Université Paris Ouest Nanterre La Défense)

7.1 Introduction

According to the Dual Mechanism approach to word processing (Prasada and Pinker 1993; Pinker and Prince 1988; Pinker and Ullman 2002; among others), recognition of a morphologically complex input word involves two interlocking steps: (i) preliminary full-form access to the lexicon, and (ii) optional morpheme-based access to sub-word constituents of the input word, resulting from the application of combinatorial rules taking care of on-line word segmentation. Algorithmically, step (ii) is taken if and only if step (i) fails to find any matching access entry in the lexicon. The view accounts for the appropriate mastering of irregular and subregular forms, which are assumed to be stored and accessed in the lexicon as full forms, while relying on morpheme-based access for the stem and affix of those morphologically

* The present work received funding from the European Community’s Seventh Framework Programme under grant agreement no FP7-231453 (HUMANOBS, Humanoids That Learn Socio-Communicative Skills Through Observation).
regular forms which are not found in the lexicon as full forms. The approach endorses a *direct functional correspondence* between principles of grammar organization supporting the lexicon vs. rules dichotomy, processing correlates (storage vs. computation), and localization of the cortical areas functionally involved in word processing (temporo-parietal vs. frontal areas in the human cortex; see Ullman 2004).

Although such a direct correspondence is probably the most straightforward model of the grammar-processing relation (Miller and Chomsky 1963; Clahsen 2006), it may only be the artefact of an outdated view of lexical *storage* as more ‘costly’ than computational operations (Baayen 2007). Alternative theoretical models put forward a more nuanced *indirect correspondence* hypothesis, based on the emergence of morphological regularities from independent principles of hierarchical organization of lexical information (Corbett and Fraser 1993; Wunderlich 1996; Dressler et al. 2006). 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). A less symbolic version of the same hypothesis (Bybee 1995a) sees the morphological lexicon as a dynamic, parallel network of fully memorized word forms. In the network, forms sharing meaning components and/or phonological structure are associatively connected with one another, as a function of formal transparency, item frequency, and size of morphological family. This view prompts a different computational metaphor than traditional rule-based models: a speaker’s lexical knowledge corresponds more to one large relational database than to a general-purpose automaton augmented with lexical storage (Blevins 2006), thus supporting a *one-route model* of word competence.

In this chapter, we explore the implications of the two models in connection with issues of inflectional paradigm learning, based on two sources of empirical evidence: (a) experimental and developmental data of human processing and storage of complex forms, with particular emphasis on dissociation effects of regular vs. irregular inflections (section 7.2) and (b) evidence of computer models of inflection learning, tested on samples of realistically distributed training data (section 7.3). We then present an original computer model of memory self-organization (section 7.4) and apply it to the task of learning verb paradigms in Italian and French (section 7.5). The model learns a stochastic finite state automaton based on patterns of Hebbian connectivity in a self-organizing topological memory. We eventually discuss the implications of this processing architecture and some experimental results against the background of the dual-route vs. one-route mélée.

### 7.2 Word processing evidence

**Morphological ontology**

In Dual Mechanism Models, roots and affixes are the basic building blocks of morphological competence, on the assumption that the lexicon is largely redundancy-free. The speaker, having identified the parts of a word form, proceeds to discard the original word from the lexicon. Contrary to such views, most one-route models take full words as basic, with sub-word constituents being considered epiphenomenal.

Over the past three decades, a large body of empirical evidence has suggested that sub-word constituents do play a crucial role in the processing and representation of morphologically complex words (see McQueen and Cutler 1998 and Clahsen 1999 for overviews). In lexical decision tasks (Taft 2004), target lexical bases are effectively primed by earlier presentation of regularly inflected related forms (*walked* → *walk*), but they are not primed by irregular inflections (e.g. *brought* vs. *bring*).

The effect is interpreted as showing that *walked* activates two distinct lexical representations, one for the stem *walk* and the other for the affix *-ed*.

Associative models of morphological processing account for dissociation effects of this kind in terms of type/token frequency factors, phonological and semantic similarity, or both (e.g. Eddington 2002; Ellis and Schmidt 1998; Joanisse and Seidenberg 1999). For example, Rueckl and Raveh (1999) argue that regular past tense forms are orthographically and phonologically more similar to their base forms than irregular past tense forms are (compare *walked* vs. *walk*; *taught* vs. *teach*); these different form properties account for full priming of regular past tense forms.

**Rule gradient**

The elicited production method allows the testing of generalization properties associated with morphological patterns. Subjects are presented with nonce-words (e.g. *tring*) for which they are asked to provide specific related inflected forms (e.g. the corresponding past participle form, say *prung*). By modulating nonce-words by similarity to attested patterns (*string–strung*) and by the frequency of these patterns (frequent vs. rare), properties of morphological processes are investigated. Productive morphological rules are reported to generalize to nonce-words irrespective of the frequency and level of similarity of attested patterns. In contrast, minor morphological processes are sensitive to such effects. Once more, this is interpreted as a memory effect.

Supporters of one-route models, on the other hand, conceive of this opposition as a gradient. More default rules may concurrently take care of
the same morphological process, possibly applied to different base forms (see Burzio’s 1998 notion of multiple correspondence). Both regular and subregular inflections typically cluster into phonologically or even semantically coherent families. Speakers demonstrably use these patterns to produce novel forms by analogy to already stored ones, and the same is true for regular inflections (see Albright’s 2002 reliability islands).

Derivationality

Dual Mechanism Models assume that base forms and fully inflected forms are derivationally related: the morphological processor accesses lexical bases to derive surface forms on-line. Alternatively, one-route models typically assume storage of full forms, both regular and irregular. On closer scrutiny, however, the derivational assumption appears to be orthogonal to the Dual Mechanism vs. one-route debate. According to some scholars (Aronoff 1994; Anderson 1982; 1992; Zwicky 1985; Carstairs-McCarthy 1987; Stump 1993; 2001), the paradigm contains a set of slots defined in terms of morpho-syntactic feature values and shows how each slot is to be filled in through application of formal functions to lexical bases. In this respect, paradigmatic relations are equivalent to augmented derivational processes, applying under some global constraints such as blocking, completeness, and uniqueness (Aronoff 1976; Carstairs-McCarthy 1987; Wunderlich 1996; Kiparsky 1998).

Frequency effects

Some important empirical findings suggest that surface word relations constitute a fundamental domain of morphological competence. Of late, particular emphasis has been laid on the interplay between form frequency, family frequency, and family size effects within morphologically based word families. The two best-known such families are the inflectional paradigm and the derivational family. Family frequency has been shown to correlate positively with response latencies in lexical decision (Baayen, Dijkstra, and Schreuder 1997; Taft 1979; Hay 2001). Family size is known to negatively correlate with visual lexicon decision latencies, as documented for a variety of languages (Baayen et al. 1997; Ford, Marslen-Wilson, and Davis, 2003; Lüdeling and Jong, 2002; Moscoso del Prado Martín, Bertram, Häkiö, Schreuder, and Baayen 2004). Evidence from research on speech errors (Stemberger and Middleton, 2003) suggests that English present and past tense forms are in competition, and that this competition is modulated by the a-priori probabilities of the vowels in these forms, even if they are regular (Tabak, Schreuder, and Baayen 2005). Finally, Maratsos (2000) reports that many individual verbs are used by children in both correct and overgeneralized forms (e.g. brought and *bringed) for a long period. The evidence seems to support a more dynamic, frequency-based competition between regular and irregular forms than dual-route accounts are prepared to acknowledge. Assuming that both regular and irregular forms are stored in the lexicon seems to go further towards a competition-based account.

Automatic morphological processing

That more than just storage is involved, however, is suggested by interference effects between false friends (or opaque pseudo-derivations) such as broth and brother, which share a conspicuous word onset but are not related morphologically (Longtin, Segui, and Mallé 2003; Rastle and Davis 2004). These and other similar results, observed particularly but not exclusively for Semitic languages (see Frost, Forster, and Deutsch 1997 and more recently Post et al. 2008), show that as soon as a given letter sequence is fully decomposable into morphological formatives, word parsing takes place automatically, prior to lexical look-up.

Paradigm learning

In the psycholinguistic literature, there is a general consensus that Italian children are more precocious in mastering the present indicative sub-paradigm than English children are in learning the simple contrast between the third singular person and the base form (Brown 1973; Pizzuto and Caselli 1992; Hlyams 1992; Noccetti 2003). Within the framework of Natural Morphology (Dressler et al. 1987), the development of verb inflection has been investigated cross-linguistically by focusing on the structural properties of morphological paradigms (Bittner, Dressler, and Kilani-Schoch 2003; Dressler 2000). Typological evidence of this kind provides a strong indication that inflectional contrasts in prototypically inflecting verb systems are acquired at a considerably earlier stage than inflectional contrasts in more isolating verb systems, in contrast with rule-based accounts of morphology learning which predict that more complex and richer (sub)paradigms should take longer to be learned.

7.3 Computational modelling

Somewhat ironically, classical multi-layered connectionist networks (see McClelland and Patterson 2002 for a review), often heralded as champions of the associative view of word structure, appear to have problems with the extensive evidence of global family size and frequency effects reported in the previous section. By modelling inflection as a phonological mapping function
from a lexical base to its range of inflected forms, connectionist architectures are closer to a sub-symbolic, neurally inspired variant of classical derivational rules than to associative models of the mental lexicon.

Lazy learning methods such as the Nearest Neighbour Algorithm (Bosch, Daelemans, and Weijters 1996) or the Analogy-based approach proposed by Pirrelli and Yvon (1999) require full storage of pre-classified word forms, and make on-line use of them with no prior or posterior organization of stored items. However, there is no explicit sense in which the system learns how to analogize new exemplars to already memorized ones, since the similarity function does not change over time and the only incremental pay-off lies in the growing quantity of information stored in the database of examples. These algorithms are good at finding analogies only if they are told where to look for them.

All the approaches mentioned above are task-oriented and supervised, since they assume that training word forms are glossed with morphological information (e.g. morpho-syntactic features or morpheme boundaries). Hence, they can replicate predefined morphological classes, but cannot discover new classes. Arguably, a better-motivated and explanatory approach should be based on the self-organization of input items into morphologically natural classes with no external supervision.

There has been a recent upsurge of interest in the use of global paradigm-based constraints to minimize the range of inflectional or derivational endings heuristically inferred from unsupervised training data (Goldsmith 2001; 2006; Gaussier 1999; Baroni 2000). Goldsmith models paradigm learning as a Minimum Description Length problem (Rissanen 1989): 'find the battery of inflectional markers forming the shortest grammar that best fits training evidence', where (i) a grammar is a set of paradigms defined as lists of inflectional markers applying to specific verb classes and (ii) the training evidence is a text corpus. The task is a top-down global optimization problem and boils down to a grammar evaluation procedure. In Goldsmith's algorithm, however, the segmentation of morphemes is kept separate from their evaluation. The two processes do not come into contact and we are left with no principled answer to the problem of the interplay between word processing and the morphological organization of the speaker’s mental lexicon. Moreover, it is hard to see how a child learning morphology can possibly be engaged in a top-down search for global minima. Finally, the algorithm tells us nothing about the way novel words are assigned to existing paradigms.

This aspect is addressed by Albright (2002), who applies the Minimal Generalization Algorithm (Pinker and Prince 1988; Albright and Hayes 2002) to the acquisition of inflectional patterns in Italian conjugation. The algorithm consists in aligning lexical entailments between inflected forms to extract from them very specific context-sensitive rules mapping one form into the other. Albright shows that rules of this kind apply quite reliably, and that their reliability score (based on the number of forms for which the mapping rule makes the right prediction) correlates with human subjects’ acceptability judgement on nonce-forms. However, Albright says very little about the type of processing architecture that could support such a rule-based conception of inflectional morphology. Moreover, it is not clear how learners can home in on the right sort of frequency counts the framework requires.

Pirrelli and colleagues (2004; 2006; Pirrelli 2007) suggest modelling the mental lexicon as a topological Self-Organizing Map (SOM; Kohonen 2001). Processing and storage in a SOM are governed by local principles of similarity between input vectors (representing unsupervised training data) and the weight vectors of the map’s processing nodes (see section 7.4 for more on this). Nonetheless, due to its topological dynamics, the map is able to develop clusters of specialized nodes which reflect global distributional patterns in the training data. This makes SOMs suitable for simulating the emergence of morphological clusters through lexical storage. However, it is difficult to see how these clusters can be used for word recognition and production.

Both associative and Dual Mechanism Models find it hard to account for the entire body of evidence reviewed here. All in all, the evidence lends support to a less deterministic and modular view of the interaction between stored word knowledge and on-line processing than dual-mechanism approaches are ready to acknowledge. If lexical blocking is assumed to transfer to word recognition, it would predict that pseudo-affixed monomorphemic words such as brother should not undergo decompositional processing, contrary to evidence on automatic processing. On the other hand, there is no way to account for such effects in terms of either variegated analogy (of the sort used by example-based approaches) or phonological complexity and perceptual sublety of the input word (as suggested by McClelland and Patterson 2002). Both analogies and inflectional rhyming patterns have to exhibit a clear morphological status; but such a status is taken to be epiphenomenal in current connectionist thinking.

Computer models have been successful in tackling certain aspects of word learning, but have not been able to provide, to date, a comprehensive picture of the complex dynamics between computation and storage underlying morphological processing. The currently emerging view sees word processing as the outcome of simultaneously activating patterns of cortical connectivity reflecting (possibly redundant) distributional regularities in the input at the phonological, morpho-syntactic, and morpho-semantic levels. At the same time, there is evidence to argue for a more complex and differentiated neuro-biological substrate for human language than connectionist one-route models.
can posit (Post et al. 2008), suggesting that brain areas devoted to language processing maximize the opportunity for using both general and specific information simultaneously (Libben 2006), rather than maximize processing efficiency and economy of storage. To our knowledge, no current computational model of word learning embodies such a complex interaction.

In what follows we describe an original computer model of dynamic memory able to simulate effects of morphological self-organization that mirror important distributional properties of inflectional paradigms. Moreover, we show that the resulting patterns of time-bound connectivity between stored items function like a stochastic processing model of word inflection that uses rule-like generalizations over learned data.

7.4 Computer modelling of memory self-organization

7.4.1 Kohonen Self-Organizing Map (KSOM)

Kohonen’s Self-Organizing Maps (or KSOMs; Kohonen 2001) are unsupervised clustering algorithms that mimic the behaviour of so-called brain maps, medium to small aggregations of neurons in the cortical area of the brain, involved in selectively processing homogeneous classes of sensory data. Processing in a brain map consists in the activation (or firing) of one or more neurons each time a particular stimulus is presented. A crucial feature of brain maps is their topological organization: nearby neurons in the map are fired by similar stimuli. Although some brain maps are taken to be genetically pre-programmed, there is evidence that at least some aspects of such global neural organization emerge as a function of the sensory experience accumulated through learning (Jenkins, Merzenich, and Ochs 1984; Kaas, Merzenich, and Killackey 1983).

A KSOM is a grid of parallel processing nodes, also suggestively referred to as 'receptors'. Each node is synaptically connected with all units on the input layer, where input vectors are encoded (Figure 7.1a). Each connection is treated as a communication channel with no time delay, whose synaptic strength is given by a weight value. Each receptor is thus associated with one synaptic weight vector in the spatial connection layer.

Weight values on the connection layer are adjusted dynamically through learning on the basis of two key principles: similarity and clustering. To see them in action, it is useful to conceive of learning as articulated into three phases: (i) parallel activation, when all receptors are fired by an input vector as a function of the similarity between their weight vector and the input vector itself; (ii) filtering, when the node whose synaptic weight vector is the most similar to the current input vector is singled out as the Best Matching Unit (BMU); and (iii) adaptive learning, when weight vectors of all receptors are adjusted to make them closer to values in the current input vector. The last step, illustrated in Figure 7.1b, is modulated by two parameters: the learning rate and the neighbourhood gain function. The learning rate defines the propensity of the map to adjust its synaptic weights. The neighbourhood function is defined as a bell-shaped curve (a Gaussian) centred on the current BMU. The further away from the BMU a node is, the lower the value on the bell and the weaker the adjustment of the node’s weight vector. Both learning rate and neighbourhood gain gradually shrink during learning, to simulate the behaviour of a map whose plasticity decreases over time.

Such a simple learning dynamics prompts an overall topological organization of the map receptors in the map space. Input vectors that are similar in the input space will strongly activate nodes that are close in the map space, as shown pictorially in Figure 7.2, where input items are assigned to three classes, each represented by a different grey pattern. On the untrained KSOM, nodes that are fired by the same class of input vectors are randomly scattered (Figure 7.2b). After training, they cluster in topologically connected areas of the map (Figures 7.2c and 7.2d).

7.4.2 Temporal Hebbian Self-Organizing Map (THSOM)

Temporal Hebbian Self-Organizing Maps (THSOMs; Koutnik 2007) model synchronization of two BMUs firing at consecutive time steps. This means
that a THSOM can remember, at time $t$, its state of activation at time $t-1$ and can make an association between the two states. This is possible by augmenting traditional KSOMs with an additional layer of synaptic connections between each single node and all other nodes on the map (Figure 7.3).

Connections are treated as communication channels whose synaptic strength is measured with a weight value, updated in a fixed one-step time delay. Weights on the connection layer (hereafter referred to as the temporal connection layer) are adjusted by Hebbian learning, based on activity synchronization of the BMU at time $t-1$ and the BMU at time $t$. During training, the temporal connection between the two BMUs is potentiated (Figure 7.4a), while the temporal connections between all other nodes and the BMU at time $t$ are depressed (Figure 7.4b). Logically, this amounts to enforcing the entailment $B_{t-1} \rightarrow B_t$.

### 7.4.3 Topological Temporal Hebbian Self-Organizing Map ($T^2$HSOM)

The model adopted in the present work originally extends Koutnik’s THSOM by using the neighbourhood function as a principle of organization of connections in the temporal connection layer (Figures 7.5a, b). An additional depressant Hebbian rule penalizes the temporal connections between the BMU at time $t-1$ and all nodes lying outside the neighbourhood of the BMU at time $t$ (Figure 7.5c). This is equivalent to the logical entailment $B_{t-1} \rightarrow B_t$. Taken together, the temporal connections in Figure 7.5 enforce a bidirectional entailment between $B_{t-1}$ and $B_t$ inducing a bias for biunique first-order Hebbian connections. We shall refer to such a bias as the association biuniqueness assumption.

### 7.4.4 $T^2$HSOMs in action

When trained on time series of input vectors, a $T^2$HSOM develops (i) a topological organization of receptors by their sensitivity to similar input vectors and (ii) a first-order time-bound correlation between BMUs activated at two consecutive time steps.
Input vectors can be similar for two independent and potentially conflicting reasons: (i) they have vector representations that are close in the input space; and (ii) they distribute similarly, i.e. they tend to be found in similar sequences. A T^2HSOM tries to optimize topological clustering according to both criteria for similarity. For any sequence of input vectors, the model creates an internal representation of the chain of BMUs fired by the sequence. It is possible to map out the corresponding chain by traversing the path of Hebbian connections leading from the BMU fired by the first input vector to the one fired by the last input vector, going through all intermediate BMUs. In this respect, the map behaves like a first-order stochastic Markov model, whose states are topological clusters of class-sensitive receptors and stochastic state transitions are represented by normalized connections.

The knowledge of a trained T^2HSOM is stored in the synaptic weights of its nodes. Understanding and evaluating the map’s learning behaviour thus requires a few post-processing steps to read off information from synaptic weights. The first step consists in calibrating the trained map by assigning a label to each node. A label is the symbol to which the node is most sensitive, that is whose input vector is closest to the node’s weight vector. Since similar input vectors activate nodes topologically close on the map, labelling reveals the topological coherence of the resulting organization (Figure 7.2d). The second post-processing step involves the temporal connection layer. Connection weights $m_{ji}$ measure the synaptic strength between two consecutively activated BMUs. They are transformed into transition probabilities by normalizing the weight matrix (by columns) and then transposing it:

$$t_{ij} = m_{ji} \cdot \frac{1}{\sum_{h=1}^{N} m_{hi}}$$

Figure 7.5 T^2HSOM’s temporal layer plasticity: (a) Long-Term Potentiation; (b,c) Long-Term Depression

7.4.5 Learning bias

Due to its temporal bias for biunique first-order Hebbian connections (section 7.4.3), the map tries to internally represent each input sequence through a dedicated chain of BMUs as shown in Table 7.1 for the input sequences ABC and 123 (left panel).

<table>
<thead>
<tr>
<th>input</th>
<th>chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>A → B → C</td>
</tr>
<tr>
<td>123</td>
<td>1 → 2 → 3</td>
</tr>
</tbody>
</table>

When different sequences share the same head but have different tails, the map creates a unique chain for the shared head and bifurcates upon the split tail, as shown in the right panel of Table 7.1. If heads differ too, the map tries to develop distinct chains by duplicating receptors that are dedicated to identical symbols (see Table 7.2, left panel). In this way, the map supplies, with space, lack of a memory order greater than one. However, if topological constraints are enforced, shared subsequences are represented through shared chains (Table 7.2, right panel).
Note that the level of entropy of the map (i.e. its degree of uncertainty) differs in the two panels of Table 7.2. In the left panel, the map can predict the two symbol chains with certainty starting from the first symbol. Thus, entropy goes down to zero. In the right panel of Table 7.2, entropy increases when the chain bifurcates, as, upon activation of ‘D’, the map is in no position to anticipate with certainty which out-going connection will be taken. Hence, specialization and memory compression increase entropy. This observation is confirmed by an analysis of the dynamic behaviour of a T²HSOM as detailed in the following section.

### 7.4.6 Time-space trade-off

As the overall topological organization of the map is the result of cooperation and competition between temporal and spatial vector similarity, its generalization capabilities crucially depend on this dynamics.

When neighbourhood functions are operating, receptors that are fired by similar input vectors tend to stick together in the map space. Large areas of receptors are recruited for frequently occurring input vectors. In particular, if the same input vector occurs in different contexts, the map tends to recruit specialized receptors that are sensitive to the specific context where the input vector is found. The more varied the distributional behaviour of an input vector, the larger the area of dedicated receptors (space allowing). These dynamics are coherent with a learning strategy that minimizes entropy over inter-node connections. Moreover, it constrains the degrees of freedom to specialize receptors, since all receptors compete for space on the map. As a result, some low-frequency input vectors may lack dedicated receptors after training. By the same token, dedicated receptors may generalize over many instances of the same input vector, gaining in generality but modelling their distributional behaviour more poorly. The main consequence of poor modelling of the time-bound distribution of input vectors is an increase in the level of entropy of the map, as more general nodes present more out-going connections. However, topological generalization is essential for a map to learn symbolic sequences whose complexity exceeds the map’s memory resources (i.e. the number of available nodes). Moreover, lack of topological organization makes it difficult for a large map to converge on learning simple tasks, as the map has no pressure to treat identical input tokens as belonging to the same type (Ferro and Pirrelli, in preparation).

### 7.5 Modelling word learning

A T²HSOM learns word forms as time-series of (phonological) symbols preceded by a start-of-word symbol (‘#’) and immediately followed by an end-of-word symbol (‘$’), as in ‘#F,A,CH,CH,O,$’ (transcribed pronunciation of Italian faccio, ‘I do’). Phonological segments are encoded through n-dimensional binary vectors specifying place and manner of articulation. In learning a word form, the map is exposed to one segment at a time, in order of appearance from left to right. Upon exposure to the end-of-word symbol ‘$’, the map resets its Hebbian connections thus losing memory of the order in which words are presented. By being trained on several sequences of this kind, a T²HSOM (i) develops internal representations of phonological symbols, (ii) links them through first-order Hebbian connections, and (iii) organizes developed representations topologically. The three steps are not taken one after the other but dynamically interact in non-trivial ways, as we shall see in the general discussion.

From a linguistic viewpoint, step (i) corresponds to learning individual phonological segments by recruiting increasingly specialized receptors. Frequent phonological segments are learned more quickly than less frequent ones. Step (ii) allows the map to develop selective paths through consecutively activated BMUs. This corresponds to learning word forms or recurrent parts of them. Once more, this is a function of the frequency with which symbol sequences are presented to the map. Finally, step (iii) combines spatial and temporal information to cluster nodes topologically. Accordingly, nodes that compete for the same symbol stick together on the map. Moreover, they tend to form sub-clusters to reflect distributionally different instances of the same symbol. For example, the phonological symbol ‘A’ in ‘#F,A,CH,CH,O,$’ will fire, if space allows, a different node than the same symbol in ‘#S,E,M,B,R,A,$’ (semsbra, ‘it seems’). In what follows we consider in some detail the implications of this strategy for learning the inflectional paradigms of a language.

In two learning sessions, we trained a 10x10 T²HSOM on present indicative forms of Italian and French verbs. For each language, the set of forms was selected according to their frequency distributions by person–number feature combinations in a reference corpus. For both experiments, the same configuration of learning parameters was chosen.

| Table 7.2. Alternative chains of BMUs activated by the same input strings |
|-----------------------------|-----------------------------|-----------------------------|
| input chain input chain input chain |
| $CD_1$ $\Rightarrow x \rightarrow C \rightarrow D \rightarrow 1$ $CD_1$ $\Rightarrow x \rightarrow C \rightarrow D \rightarrow 1$ |
| $\beta CD_2$ $\Rightarrow \beta \rightarrow C \rightarrow D \rightarrow 2$ $\beta CD_2$ $\Rightarrow \beta \rightarrow C \rightarrow D \rightarrow 2$ |
7.5.1 Italian
The Italian training dataset contains present indicative forms of 20 different verbs, for a total of 103 attested different forms, whose frequency distributions are sampled from the Calambrone section of the Italian CHILDES sub-corpus (MacWhinney 2000), of about 110,000 token words. As we were mainly interested in effects of global paradigm-based organization, forms were mostly selected from regular, formally transparent paradigms. Nonetheless, some subregular high-frequency forms (such as those of STARE ‘stay’, FARE ‘make’ and POTERE ‘be able’) were present in the training set.

The resulting overall map is shown in Figure 7.6. Shades of grey on arrow lines indicate levels of connection strength, with darker lines representing stronger connections. Vowel segments are clustered topologically and clearly separated from consonants. Moreover, they appear to play the role of prominent attractors for patterns of connectivity, particularly when they function as (parts of) inflectional endings. In some cases, it is possible to follow a continuous path of connections going from ‘#’ (top left corner of the map), to ‘$’ (bottom left corner of the map), as with the high-frequency word form ‘#,F,A,$’ (‘it does’). In the vast majority of cases, however, connection chains represent recurrent segmental subsequences, corresponding to inflectional endings, verb stems, or parts of them.

The underlying paradigmatic structure of such an entangled bundle of connections is thrown into sharper relief in Figure 7.7, where each panel shows the activation pattern of the map when it is fired by forms that occupy the same present indicative cell (1st person singular, 2nd person singular, etc.). Clearly, the six panels share a substantial number of connectivity patterns, due to repeated activation of regular stems, and differ in the way stems are connected with inflectional endings.

Figures 7.6 and 7.7 provide a static view of paradigms as entrenched patterns of inter-node connectivity. To get a flavour of the process of dynamic
emergence of any such pattern through learning we have to turn to Figure 7.8. Here, temporal weights over connections between the verb stem and its present indicative endings are monitored through 100 learning epochs for the regular verb ARRIVARE (‘arrive’). In the plot, each connection weight is normalized according to equation (i) above. After an initial chaotic phase, with inflectional endings competing for primacy, the paradigm converges to a stable state at around epoch 45. The two greyish vertical lines on the plot mark the points in time when the topological organization of the spatial layer subsides (solid line), and when the topological organization of the temporal layer comes to an end (dashed line). We shall comment on the role of these two signposts later in the general discussion.

7.5.2 French

The French training set includes 100 present indicative forms sampled from the MorPa corpus (Montermini, Boyé, and Tseng 2008) according to their frequency distributions and phonologically transcribed with an inventory of 40 phonemes with binary vector encoding.

The four panels of Figure 7.9 show patterns of connections for 1st singular forms (a), 1st plural forms (b), 2nd plural forms (c) and 3rd plural forms (d) of the French present indicative. In panels (b) and (c), 1st and 2nd plural endings show dedicated patterns of connectivity reflecting their recurrent activation. Each such pattern includes the incoming connection to the node specialized for the inflectional ending (represented by the symbol ‘$’ for the first person plural and ‘e’ for the second person plural) and its outgoing connection to the end-of-word symbol ‘$’.

Figure 7.10 shows how temporal weights over stem-ending connections evolve through learning epochs in the regular present indicative paradigm of AIMER. Due to the extensive syncretism of French present indicative forms in the three singular persons (and, in regular cases, also in the third plural person), the cumulative effect of their pressure on first and second plural endings means that the latter hardly reach perceivable levels of activation.
7.6 General discussion

T\textsuperscript{2}HSOMs memorize word forms by topologically organizing (a) nodes that are sensitive to word segments and (b) patterns of time-bound connectivity between consecutively fired nodes. Due to the biuniqueness association assumption (section 7.4.3), the strength of any inter-node connection \textquote{\textquoteright}A\textrightarrow \textquoteright B\textquoteright is (a) a direct function of the number of times the connection is activated during training, (b) an inverse function of the number of times \textquoteright}A\textquoteright is seen preceding a symbol which is not \textquoteright}B\textquoteright, and (c) an inverse function of the number of times \textquoteright}B\textquoteright is seen following a symbol which is not \textquoteright}A\textquoteright. This simple dynamics has important consequences for the way paradigms are learned and eventually organized by a T\textsuperscript{2}HSOM.

First, all regularly inflected forms belonging to the same paradigm compete with one another on the map. For instance, \texttt{\textquoteright}A,R,R,Y,O,S\textquoteright\textquoteright \textquoteright}I arrive\textquoteright\textquoteright inhibits and is inhibited by \texttt{\textquoteright}A,R,R,Y,A,S\textquoteright\textquoteright \textquoteright}(s)he arrives\textquoteright\textquoteright. Second, each form may be supported by other word forms sharing the same stem-ending transition. For example, \texttt{\textquoteright}A,R,R,Y,O,S\textquoteright\textquoteright is strengthened by \texttt{\textquoteright}A,D,E,Y,O,S\textquoteright\textquoteright. Finally, \texttt{\textquoteright}A,R,R,Y,O,S\textquoteright\textquoteright is inhibited by word forms with a different stem-ending transition, such as \texttt{\textquoteright}A,S,T,O,S\textquoteright\textquoteright \textquoteright}I stay\textquoteright\textquoteright and \texttt{\textquoteright}A,L,E,G,O,S\textquoteright\textquoteright \textquoteright}I read\textquoteright\textquoteright.

To illustrate this point in more detail, let us turn back to Figure 7.8 above. In the paradigm of ARRIVARE, different endings compete chaotically in the first learning epochs, before the map reaches a stage where the topological organization of the spatial layer subsides (solid vertical line). In this initial phase, dedicated receptors have not yet developed (especially for low-frequency symbols) and BMUs change dramatically, affecting large neighbouring areas on the map. An intermediate phase starts with the solid line and ends with the dashed vertical line, marking the epoch where the topological organization of the temporal layer comes to an end. In this phase, macro-paradigms start to set in. Dedicated receptors are already topologically organized and entrenched but temporal connections are still changed over neighbouring areas. This means that different connections going into the same receptor, say \texttt{V\textrightarrow O\textquoteright} and \texttt{T\textrightarrow O\textquoteright}, strongly compete with one another for the whole cluster of \texttt{O\textquoteright} receptors. This causes connection weights to vary considerably. The higher the productivity of an inflectional ending the more chaotic this phase. Finally, when the dashed vertical line is reached, temporal weights change moderately, with excitatory connections acting locally and inhibitory connections making receptors specialize for context-sensitive symbols. All in all, this phase can be interpreted as a process of paradigm refinement, where the map assigns relative association strengths to endings that form part of the same paradigm.

![Figure 7.11](image-url) Intra-paradigmatic association strength in Italian and French

This dynamics sheds light on frequency effects of paradigm entrenchment. Figure 7.11 shows average levels of normalized association weights over stem-ending connections in the present indicative paradigms of Italian (panel a) and French (panel b), plotted against an incrementally growing lexicon. In Italian, association weights of high-frequency endings start high in the upper part of the panel, but slump rather quickly as the lexicon grows. Such a trend is counterbalanced by the characteristically U-shaped curve of weights for low-frequency endings in the same panel. After the map is exposed to 100 word forms, degrees of association strength level out considerably, allowing the map to settle down far away from its associational biuniqueness bias. The probability mass that the map assigns, on average, to an Italian present indicative paradigm tends to be more evenly distributed after training, thus avoiding within-paradigm levelling effects. This is due to balanced competition among intra-paradigmatic endings and lack of syncretism. In mathematical terms, Italian present indicative paradigms are highly entropic, and this causes fast convergence of transition probabilities in the map.

The result highlights two further points. As more words are learned, association strengths get smaller, since the map is storing more information on time-bound connections between segments. Since the map must take stock of more and more outgoing transitions from each node, paradigm entropy increases. Nonetheless, the map is memorizing word forms better, as witnessed by decreasing levels of the map's uncertainty in going through a known
word (Figure 7.12). Hence paradigm entropy increases, while word entropy decreases. The map is moving towards a more balanced organization of verb forms into paradigms, while, at the same time, memorizing individual verb forms increasingly more accurately.

The Italian overall trend is less prominent when we look at the French data (Figure 7.11b). Association weights neither go down nor converge as quickly as in the Italian experiment. There is a slowly decreasing trend in the overall association strength, but differences over individual endings remain high. After being exposed to 100 different word forms, the map finds it difficult to distribute transition probabilities evenly within a paradigm. The extensive syncretism of French present indicative forms produces a frequency gang effect that slows down the process of learning less frequent inflectional markers (Figure 7.10).

### 7.7 Concluding remarks

As descriptive tools of theoretical linguistics, paradigms have enjoyed a hybrid status, halfway between entrenched patterns of lexical organization and processing structures enforcing global constraints on the output of traditional inflection rules. In a psycholinguistic perspective, they appear to play a significant role not only in the way morphological information is processed, but also in the way the same information is acquired and structured through competition of concurrently memorized word forms. To our knowledge, no existing computational or psycholinguistic model of morphological processing can capture such a manifold range of diverse and potentially conflicting requirements.

In the present contribution, we show that paradigmatic structures can emerge through word learning as the by-product of the endogenous dynamics of lexical memorization as competitive self-organization, based on the diverging principles of formal contrast (in space) and association biuniqueness (in time). According to this view, inflected forms are not the output of rules mapping lexical representations onto surface realizations, but rather the driving force of lexical organization. However, more than storage is involved here. A trained map behaves like a first order stochastic Markov chain, with inter-node connections building expectations about possible inflected forms on the basis of a global topological organization of already known forms.

The model, we contend, prompts a radical reappraisal of the traditional mélée between one-route and dual-route models of morphology processing and learning. That patterns of morphological structure are derivative of associative connections between stored forms (modulated by frequency) by no means implies that the same patterns play no role in word processing. Being derivative does not necessarily mean being epiphenomenal. Conversely, that rule-like generalizations apply in an apparently context-free way does not imply that they play no role in the way word information is structured and organized in the lexicon. Being important to processing does not mean being irrelevant for word learning and storage, and vice versa. We believe that further investigation into the computational and neuro-biological substrates of morphological paradigms is certain to change our views on foundational issues of grammar architecture.
References


Garcia Valdés, César Carmen (1979). El habla de Santanans de Pravia. Mieres del Camino: Instituto "Bernaldo de Quirós".


Pace, Anna (1994). Ricerche di morfosintassi sui dialetti di Trebisacce e Castrovillari. Tesi di laurea, Università della Calabria.


Pirrelli, Vito, Calderone, Basilio, Herreros, Ivan and Virgilio, Michele (2004). ‘Non-locality all the way through: Emerging global constraints in the Italian morphological lexicon’, Proceedings of 7th Meeting of the ACL SIGPHON. Barcelona: ACL.


Russell, Kevin (1995). 'Morphemes and candidates in Optimality Theory', Ms., Rutgers Optimality Archive, version X.0.s.95. [http://roa.rutgers.edu/files/44-0195/44-0195-RUSSELL-0.0.PDF].

Russi, Cinzia (2005). 'Semantic change and pragmatic inference in the grammaticalization of Italian Id. Paper given at the 6th High Desert Linguistics Conference, University of New Mexico, Albuquerque, 4–6 November.


References


Santos, Maria José de Moura (1967). 'Os falares fronteiriços de Trás-os-Montes', Revista portuguesa de filologia 12.2: 12–14.


Vasconcellos, José Leite de (1900). Estudos de filologia mirandesa. Lisboa: Imprensa nacional.


Index

Abellé A. 401, 415
Abruzzese 353
absolute tense 421, 433
abstracted paradigm model 401, 410–414, 416
Acadian French see French, Acadian
Ackerman, F. Y. 401, 403, 405, 406, 408, 419
Acquaviva 360m
acquisition 31, 140, 228, 230, 243, 244, 279, 281, 283, 307–308/84, 324, 434
Adams, J. N. 428
adjective 29, 30, 33, 40, 41, 44, 54, 61, 97, 98, 104, 109, 114, 116, 117, 175, 197, 205, 212, 244, 249, 329/93, 332, 333, 336, 339, 341, 343, 347–350, 358/626, 368, 407, 429, 432
adstrate 237, 244/6
affectedness 387, 398
affix ordering 236
agglutinating 426
Agononese 331, 339, 340, 350, 354
agreement 1, 8, 25, 82, 83, 81/86, 84, 135, 237, 239, 242–244, 253, 294, 327–357, 360/631, 391, 403, 405, 408, 427, 428
Air 302
Aissen, J. 242
Alighiero 208
Align 55, 59–65, 68, 125, 158
allotropy 359
Altamura 131, 348, 348/109, 350
analogue change 74
analogue levelling 126, 127, 129, 131, 283, 308
analogue remodelling 129, 131, 314–315/96
analogue spread 90, 93
analogy 384/9, 57, 68, 88, 102, 111, 123, 138, 141, 179, 189/95, 190, 197, 199/94, 201–203, 211, 212, 226, 227, 311, 312, 313/35, 314, 316, 318/77, 318/80, 322, 324–325, 399
anaphora 387, 422/4
Andalusian Spanish see Spanish, Andalusian
Andersen, H. 13
Andrade, A. 160
Anglo-Irish 308
Angular 244
animate 348/19
aorist 359, 259/91, 422/95
Arabic 260, 262, 263, 267, 270
Arregi, K. 52, 54, 82, 83, 83/99, 83/100
article 5, 7, 9, 31, 65, 125, 137/14, 210, 337–342, 342/17, 344, 347–351
aspect 3, 4, 20, 26, 49, 55, 70, 91–98, 140–142, 147, 191, 197, 199, 209, 236, 266, 290, 311, 324, 327, 393, 403, 405, 408, 417–417/2, 420, 421, 426, 430, 434
assimilation 21
Asturian 88, 90, 339/15
Athapaskan 258
Atlantic creoles 241–242
attraction 120, 121, 129, 194, 208
Augure 290, 296/12, 297
augment 31, 42–46
Aulus Gellius 427/13
Aurnis 312
Austin, P. 172, 382
autonomy 1–3, 31/17, 4, 6, 63, 137, 179, 235, 236, 418, 426, 430, 431
Averera, J. 123/17, 133
auxiliary 70, 311, 357, 403–409, 418, 419, 422, 425, 428, 430
Avelo, F. 171
Badietto 356, 357
Baie des Chaleurs 312