Prediction and Generalisation in Word Processing and Storage

Claudia Marzi
Institute for Computational Linguistics"A. Zampolli"- CNR Pisa,
University of Pavia - Dept. of Theoretical and Applied Linguistics
claudia.marzi@ilc.cnr.it

Marcello Ferro
Institute for Computational Linguistics"A. Zampolli"- CNR Pisa
marcello.ferro@ilc.cnr.it

Vito Pirrelli
Institute for Computational Linguistics"A. Zampolli"- CNR Pisa
vito.pirrelli@ilc.cnr.it

1. Introduction

Word storage and processing have traditionally been modelled according to different computational paradigms, in line with the classical corner-stone of "dual-route" models of word structure assuming a sharp dissociation between memory and computation (Clahsen 1999, Di Sciullo & Williams 1987, Pinker & Prince 1988, Parasada & Pinker 1993). Even the most radical alternative to dual-route thinking, connectionist one-route models, challenged the lexicon-grammar dualism only by providing a neurally-inspired mirror image of classical base-to-inflection rules, while largely neglecting issues of lexical storage (Rumelhart & McClelland 1986, McClelland & Patterson 2002, Seidenberg & McClelland 1989). Recent psycho- and neuro-linguistic evidence, however, supports a less deterministic and modular view of the interaction between stored word knowledge and on-line processing [Baayen et al. 1997, Hay 2001, Maratsos 2000, Stemberger & Middleton 2003, Tabak et al. 2005, Ford et al. 2003, Post et al. 2008). The view entails simultaneous activation of distributed patterns of cortical connectivity encoding redundant distributional regularities in language data. Furthermore, recent developments in morphological theorising question the primacy of grammar rules over lexical storage, arguing that word regularities emerge from independent principles of lexical organisation, whereby lexical units and constructions are redundantly stored and mutually related through entailment relations (Matthews 1991, Corbett & Fraser 1993, Pirrelli 2000, Burzio 2004, Booij 2010). We endorse here such a non modular view on Morphology to investigate two basic behavioural aspects of human word processing: morphological prediction and generalisation. The investigation is based on a computer model of morphology acquisition supporting the hypothesis that they both derive from a common pool of principles of lexical organisation.

2. Background

2.1. Generalisation

Morphological generalisation is at the roots of the human ability to develop expectations about novel lexical forms, so that some words (say *plipped* as the past tense of the nonce verb *plip*) are perceived by speakers as more acceptable than other potential competitors (e.g. *plup* or *plept* for the same base). These expectations can be used to produce novel forms from familiar bases. After Berko's seminal work on children mastering *wug* words (1958), linguists have put considerable effort into trying to unravel conditions for generalisations in the morphological competence of both learners and mature speakers (Bybee and Pardo, 1981; Bybee and Slobin, 1982; Bybee and Moder, 1983). After the advent of connectionism (Rumelhart and McClelland, 1986), the question of what

structural and formal conditions affect morphological generalisations in humans was coupled with the substantially different question of whether artificial neural networks eliminate or rather sub-symbolically implement algebraic productive rules of some kind (Smolensky 1988). The ensuing debate went through controversial issues of grammar architecture, centred on the hypothesis of a sharp separation between lexicon (functionally related to storage) and rules (functionally related to processing). A recent reformulation of the problem of morphological generalisation is due to Albright and Hayes (2003): given that many morphological processes are known to be productive in limited contexts, what sort of computational mechanisms are needed to account for context-sensitive restrictions on morphological generalisations? Albright and Hayes suggest that speakers conservatively develop structure-based rules of mapping between fully-inflected forms. In the orthographic domain, a mapping pattern such as X*s* → X*ing* accounts for the word pairs *talks-talking*, *plays-playing*, *forms-forming* etc., but would wrongly yield *puts-*puting* and *gives-*giveing*. These patterns are based on a cautious inductive procedure named "minimal generalisation", according to which speakers are confident in extending a morphological pattern to other forms to the extent that i) the pattern obtains for many existing word forms and ii) there is a context-sensitive difference between those word forms and word forms that take other patterns. For example, a speaker has to induce the more specific pattern X*ts* → X*tting* to cover *puts-putting*, *sits-sitting*, *hits-hitting* etc. An important point made by Albright and Hayes is that patterns apply to similar word pairs, with word similarity being based on a context-sensitive structural mapping', rather than on a pre-theoretical notion of "variegated" analogy. Finally, the level of confidence of a speaker in the pattern is defined by the ratio between the number of forms undergoing the pattern change and the number of forms meeting the context for the pattern change to apply.

A number of interesting theoretical implications follow from Albright and Hayes' approach. Unlike traditional dual-route models of morphological competence, their minimally generalised patterns are not committed to a derivational conception of morphological generalisation, according to which rules define base-to-form mapping relations only. Patterns may underlie any pair of intra-paradigmatically related forms. This view is easily amenable to a word-and-paradigm conception of the morphological lexicon where fully inflected forms are redundantly stored and mutually related through entailment relations (Matthews 1991; Pirrelli 2000; Burzio 2004; Blevins 2006). Accordingly, the speaker's knowledge of word structure is more akin to one dynamic relational database than to a general-purpose automaton augmented with lexical storage. Nonetheless, mapping patterns adhere to a rule-like manner of stating generalisations, providing the necessary and sufficient conditions that a form must meet in order for the pattern to apply. Albright and Hayes argue that this does not have to be true for variegated analogy to apply, as, in principle, forms undergoing the same pattern change may be similar to one another in many different ways, thus going beyond the reach of a structural rule-like description of the needed context. Finally, they claim that sensitivity to context-based similarity is not a specific condition of unproductive morphological processes (as suggested by dual-route modellers), but a hallmark of any morphological pattern change. Cautious generalisation is an inherent feature of morphological productivity as such.

2.2. Prediction

Morphological prediction defines the human capacity to anticipate upcoming known words. Unlike generalisation, which refers to the ability to go beyond available evidence and compensate for gaps in lexical competence, prediction appears to functionally maximise available linguistic evidence (including, but not limited to, lexical competence) to entertain hypotheses about the upcoming flow of language input, and make language

comprehension easier and more efficient. From a more general perspective, experimental studies based on event-related potentials and eye-movement evidence, for example, show that people use prior (lexical and semantic) contextual knowledge to anticipate upcoming words (Altmann and Kamide 1999; Federmeier 2007). DeLong et al. (2005) demonstrate that expected words are pre-activated in the brain in a graded fashion, reflecting their expected probability. This provides the empirical ground to probabilistic approaches to lexical prediction and gaze planning in reading. Ferro et al. (2010) offer a computational model of the interlocked relationship between processes of lexical self-organisation and active sensing strategies for reading that exploit expectations on stored lexical representations to drive gaze planning. This can explain why the capacity to repeat non words is a good predictor of whether or not the child is likely to encounter reading problems (Baddeley and Gathercole 1992; Gathercole and Pickering 2001).

There have been attempts to explain the role of prediction in facilitating language comprehension on the basis of the argument that highly predictable words are easier to integrate into the linguistic context (e.g. because unexpected words in test experiments often violate the grammatical constraints imposed by the context itself). In fact, recent evidence goes against this simpler explanation, suggesting that predictions can be made at many different levels of language comprehension, including strong biases against perfectly grammatical but somewhat rarer or less likely lexical alternatives (Staub & Clifton 2006, DeLong et al. 2005). A more intriguing explanation comes from evidence of mirror neurons (Wilson & Knoblich 2005) pointing to the observation that perceiving other people's behaviour activates covertly imitative motor plans. The use of covert imitation to facilitate perception of other's people behaviour could occur in any domain where upcoming behaviour is at least sometime predictable and where the perceiver can also perform that behaviour. Language in general, and lexical access in particular, are cases in point.

2.3. Grammar, Memory and the Lexicon

At its core, the lexicon is the store of words in long-term memory. Any attempt at modelling lexical competence must take into account issues of string storage. 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 had the unquestionable merit of highlighting some fundamental issues of coding, maintenance and manipulation of strings of symbols. It is somewhat surprising that the linguistic literature on lexical access and organisation, on the one hand, and the psycho-cognitive literature on memory processes on the other hand, have so far made comparatively sparse contact. This is arguably due to the strong influence of the calculator metaphor (Baayen 2007) on mainstream conceptions of the role of the lexicon in the grammar architecture. According to the metaphor, the lexicon is only storage, an inert repository of item-based, unpredictable information whose nature and structure is predetermined, and considered as relatively unproblematic. The combinatorial potential of lexical items, on the other hand, is defined by the rules of grammar, taking care of processing issues. Contrary to what is commonly held, connectionism has failed to offer an alternative view of such an interplay between storage and processing. There is no place for the lexicon in classical connectionist networks: in this respect, they seem to have adhered to a cornerstone of the rule-based approach to morphological inflection, thus providing a neurally-inspired mirror image of derivational rules.

In this paper, we entertain the substantially different view that memory plays a fundamental role in lexical modelling, and that computer simulations of memory processes can go a long way in addressing issues of lexical acquisition and processing. Recent studies of cortico-cortical evoked potentials show a functional bidirectional

connectivity between anterior and posterior language areas (Matsumoto et al., 2004), pointing to more integrated and dynamic mechanisms underlying language functioning in the brain than previously acknowledged. In addition to its well-established linguistic functions, Broca's area appears to be engaged in several other cognitive domains such as music (Maess et al., 2001), working memory and calculation, as well as action execution and understanding (Buccino et al., 2001; Fadiga et al., 2009). Taken together these results suggest the intriguing possibility that Broca's areas could provide the neural structures subserving context-dependent sequence processing in general, and that these structures shed considerable light on what we know about lexical organisation, access and productivity.

Human lexical competence is known to require the fundamental ability to retain sequences of items (e.g. letters, syllables, morphemes or words) in the working memory (Gathercole and Baddeley, 1989; Papagno et al., 1991). Speakers appear to be sensitive to frequency effects in the presentation of temporal sequences of verbal stimuli. Items that are frequently sequenced together are stored in the Long-Term (LT) Memory as single chunks, and accessed and executed as though they had no internal structure. This increases fluency and eases comprehension. Moreover, it also explains the possibility to retain longer sequences in Short-Term (ST) Memory when familiar chunks are presented. The ST span is understood to consist of only a limited number (ranging from 3 to 5 according to recent estimates, e.g. Cowan 2001) of available store units. A memory chunk takes one store unit of the ST span irrespectively of length, thus leaving more room for longer sequences to be temporarily retained. Furthermore, chunking produces levels of hierarchical organisation of the input stream: what is perceived as a temporal sequence of items at one level, may be perceived as a single unit on a higher level, to become part of more complex sequences (Hay and Baayen 2003). Finally, parts belonging to high-frequency chunks tend to resist being perceived as autonomous elements in their own right and being used independently. As a further implication of this "wholeness" effect, frequently used chunks do not participate in larger word families (e.g. inflectional paradigms).

From this perspective, generalisation and prediction can be seen as being in competition. Prediction presupposes LT entrenchment of memory chunks as a result of repeated exposure to frequent sequences of letters/segments. LT entrenchment eventually drives word recognition through anticipatory activation of frequently-activated chunks. Prediction is thus most accurate when concurrent activation of LT chunks is minimised. In information theoretic terms, this is equivalent to minimising the entropy over lexical choices. Generalisation, on the other hand, requires that the lexicon contains recurrent sub-lexical chunks, which recombine for novel words to be recognised as well-formed. This is equivalent to keeping entropy high in the lexicon, making room for novel word stimuli.

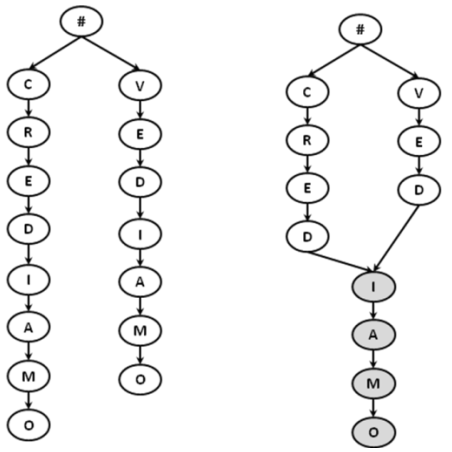

Figure 1: A word-trie (left) and a word-graph (right), for the Italian forms *VEDIAMO* ('we see') and *CREDIAMO* ('we believe').

Figure 1 shows examples of lexical structures that can account for these effects. So-called "word-tries" (left) encode symbol sequences as rooted hierarchies of labelled nodes connected through arcs, under the constraint that no node can be reached by two different descending arcs. So-called "word-graphs" (right), on the other hand, allow the same node to be reached by multiple arcs, thus using up fewer nodes in representing partially overlapping forms. For our present concerns, word-tries can be seen as encoding deeply entrenched, dedicated memory structures, whereby partially overlapping forms are nevertheless assigned independent representational resources. On the other hand, word-graphs allow for shared substrings to be assigned identical memory units. As shown below, the two graph types can in fact be conceived of as different developmental stages in lexical acquisition. In the ensuing sections, we offer a computational model of dynamic memories that can explain the emergence of such lexical structures in terms of common computational principles of self-organisation and time-bound prediction.

3. Hebbian SOMs

Kohonen's Self-Organizing Maps (*SOMs*) (Kohonen, 2001) define a class of unsupervised artificial neural networks that mimics the behaviour of small aggregations of neurons (*pools*) in the cortical areas involved in the classification of sensory data (*brain maps*). In such aggregations, processing consists in the activation of specific neurons upon presentation of a particular stimulus. A distinguishing feature of brain maps is their topological organisation (Penfield and Roberts, 1959): nearby neurons in the map are activated by similar stimuli. Although some brain maps can be pre-determined genetically, there is evidence that at least some aspects of their neural connectivity emerge through self-organisation as a function of cumulated sensory experience (Kaas et al., 1983). Functionally, brain maps are thus dynamic memory stores, directly involved in input processing, exhibiting effects of dedicated long-term topological organisation.

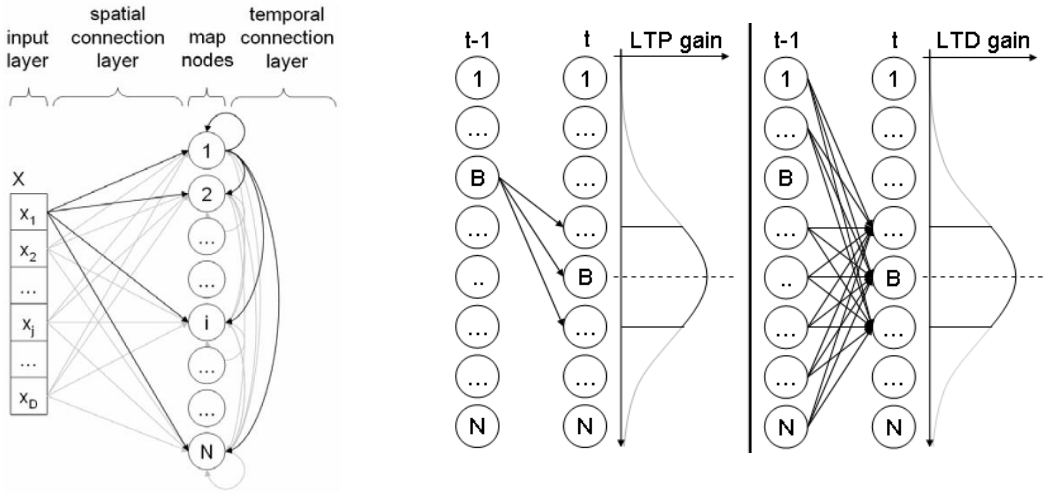

Figure 2: Left: Outline architecture of a T2HSOM. Each node in the map is connected with all nodes of the input layer. Each connection is a communication channel with no time delay, whose synaptic strength is modified through training. Connections on the temporal layer are updated with a fixed one-step time delay, based on activity synchronisation between *BMU(t–1)* and *BMU(t)*. **Right**: Long Term Potentiation (LTP) and Long-Term Depression (LTD) of Hebbian connections between consecutively activated nodes in the learning phase. (from Ferro et al., 2010)

In its typical configuration (Kohonen, 2001), a *SOM* is a grid of parallel processing nodes fully connected to an *input layer* where incoming stimuli are encoded. Input connections are modelled as weighted communication channels with no time delay, defining a spatial layer (SL) of connectivity. In the present work we make use of Topological Temporal Hebbian Self-Organizing Maps (*T2HSOMs*) (Koutnik, 2007; Ferro et al., 2010, 2011), an extension of traditional SOMs augmented with re-entrant Hebbian connections defined over a temporal layer (TL), encoding probabilistic expectations of time series. Each map node is linked to all other nodes through a delayed connection that provides, at time *t*, the activity of all nodes at time *t-1* (Fig. 2, left).

Nodes exhibit a short-term dynamic, based on equation (1) below, and a long-term dynamic, based on adaptive learning (Ferro et al., 2010). Upon presentation of an input stimulus at time *t*, all map nodes are activated synchronously at different levels $h_i(t)$, but only the most highly activated one, called the Best Matching Unit (or *BMU(t)*), is selected. Node that the activation equation (1) of node n_i at time *t* is the sum of two functions. The first function, $h_{S,i}(t)$, measures how similar the input vector weights of node n_i are to the current input, and the second one, $h_{T,i}(t)$, how predictable the current input is on the basis of past input. Parameters α and β (in 1) determine the relative contribution of the two functions to the overall activation score; high values of α make the map sensitive to the specific content of the current input stimulus, while high values of β make the map sensitive to its timing.

(1) $$BMU(t) = n_i = \arg\max_{i=1,\dots N} \{ h_i(t) \}$$

$$\text{where } h_i(t) = \alpha \cdot h_{S,i}(t) + \beta \cdot h_{T,i}(t)$$

In the learning phase, at each time *t*, *BMU(t)* adjusts its connection weights on both layers (SL and TL) and propagates adjustment to neighbouring nodes. On SL, adjustment makes connection weights closer to values in the input vector. On TL, adjustment of Hebbian connections i) potentiates the strength of association from *BMU(t-1)* to *BMU(t)* (and its neighbouring nodes), and ii) depresses the strength of association from all other nodes to *BMU(t)* (and its neighbouring nodes) (Fig. 2, right). This amounts to logically enforcing the entailment *BMU(t)* → *BMU(t-1)*, thereby inducing the emergence of dedicated patterns of activation over nodes that are reminiscent of word graphs (Fig. 1, left).

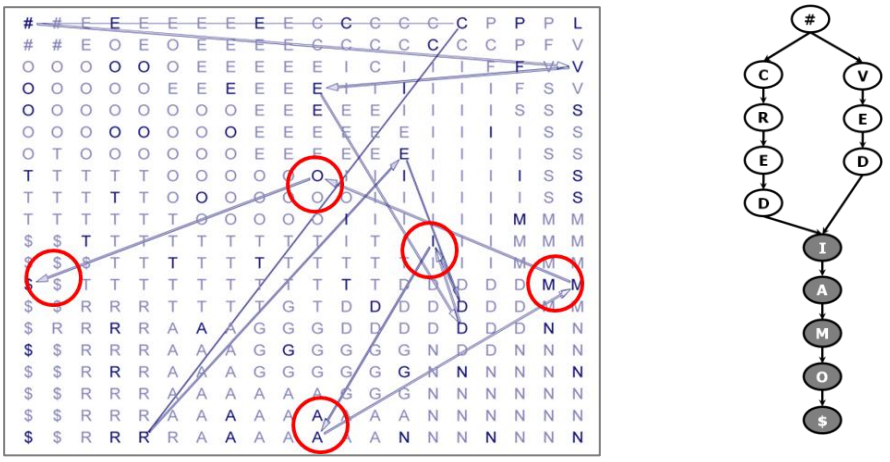

Figure 3: BMU activation chains for *vediamo-crediamo* on a 20×20 map (left) and their word-graph representation (right). Shared processing nodes are circled on the map and shaded in grey in the word graph.

When a string of letters is presented to the map one character at a time, a temporal chain of *BMU*s is activated. Figure 3 illustrates two such temporal chains, triggered by the

Italian verb forms *crediamo* and *vediamo* ('we believe' and 'we see') presented to a 20×20 nodes map trained on 30 verb paradigms, sampled from the Italian Treebank corpus (Montemagni et al., 2003) by decreasing values of cumulative paradigm frequency. In the figure, each node is labelled with the letter the node is most sensitive to. Pointed arrows represent temporal connections linking two consecutively activated nodes, thus depicting the temporal sequence of node activation, starting from the beginning-of-word symbol '#' (anchored in the top left corner of the map) and ending to '$'. Activation chains allow us to inspect the memory patterns that a map develops through training.

Although temporal learning is based on first-order re-entrant Hebbian connections only (i.e. connections emanating from the immediately preceding *BMU*), nodes can propagate information of their immediate left-context over longer activation patterns, thereby simulating orders of memory longer than 1. In Figure 3, both letters *D* in *VEDIAMO* and *CREDIAMO* are preceded by *E*. Nonetheless they recruit two topologically close but distinct nodes on the map which thus "store" information that the two *E*s were in turn preceded by a different symbol (second-order memory). A Hebbian map can thus enforce longer orders of memory through a profligate use of dedicated nodes, trading space for time. The trade-off is based on learning, and depends on available memory resources and distribution of training data. Upon hitting the ensuing *I* in *VEDIAMO* and *CREDIAMO*, the map in Figure 3 recruits the same *BMU*, showing that the map cannot retain higher-order memory events (at least for this specific sequence). The distance on the map between two *BMU*s that respond to identical symbols in different input contexts thus reflects the extent to which the map perceives them as similar. By the same token, the topological distance between chains of activated *BMU*s responding to similar input strings tells us how well the map is aligning the two strings. This is a general problem for morphology induction, arising whenever known symbol patterns are presented in novel arrangements, as when speakers are able to spot the Arabic verb root shared by *kataba* ("he wrote") and *yaktubu* ("he writes"), or the German verb root common to *machen* ("make") and *gemacht* ("made" past participle).

3.1. Inductive Bias and Input recoding

It is useful at this stage to focus on the inductive bias of Hebbian SOMs under specific parameter configurations and training conditions. Figure 4 (right) shows the topological configuration of a map trained on a uniformly distributed data set of 64 binary strings.

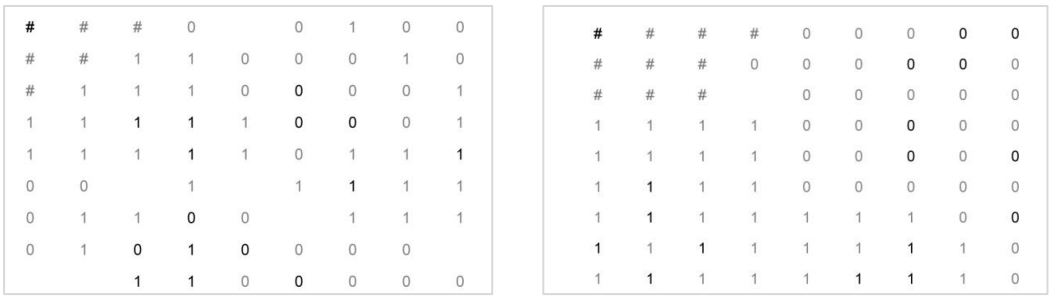

Figure 4: Topologies of a *T*-map (left) and a *ST*-map (right) trained on binary strings. Only the top left corner of both maps is shown.

The map (henceforth referred to as a Spatio-temporal map or *ST*-map for short) presents a comparatively high value for α, causing equation (1) to be more sensitive to symbol coding than to symbol timing. Conversely, the leftt-hand map of Figure 4 (called Temporal map or *T*-map for short), is more sensitive to weights on the temporal connection layer due to lower α. The resulting topologies of the two maps are different. Symbol coding defines the most external level of clustering for the *ST*-map, with all nodes

fired by a specific bit being clustered in the same connected area. In fact, due to the uniform distribution the training data, '0' and '1' take the top right corner and bottom left corner of the map respectively, parting the map's topological space into two halves. Within each half, several nodes are recruited for different instances of the same bit, as a function of their position in the training sequences. The *T*-map reverses the clustering hierarchy, with timing defining the most external level of nesting (Fig. 4, left). Within each such external cluster, nodes are specialised for being sensitive to different, similarly-distributed bits. Hence, '0' and '1' appear to be scattered around the map's topological space, depending on their time-bound distribution in the training corpus.

Differences in topological organisation define the way the two map types categorise input symbols (and eventually input sequences). A *T*-map recodes symbols positionally, by recruiting nodes that are sensitive to – say – a *1* in first position (1_1), a *1* in second position (1_2), a *0* in first position (0_1) and so on and so forth. An *ST*-map tends to categorise instances of the same symbol across different positions in the input. Nonetheless, since *ST*-maps are also sensitive to timing, they are able to distinguish instances of the same symbol on the basis of its left context. For example, a *1* preceded by a *0* will activate a dedicated 0_*1* node, whereas a *1* preceded by another *1* will activate a different 1_*1* node in the same cluster. The difference is shown in Figure 5, plotting the topological dispersion of map nodes by position of input bits for two maps.

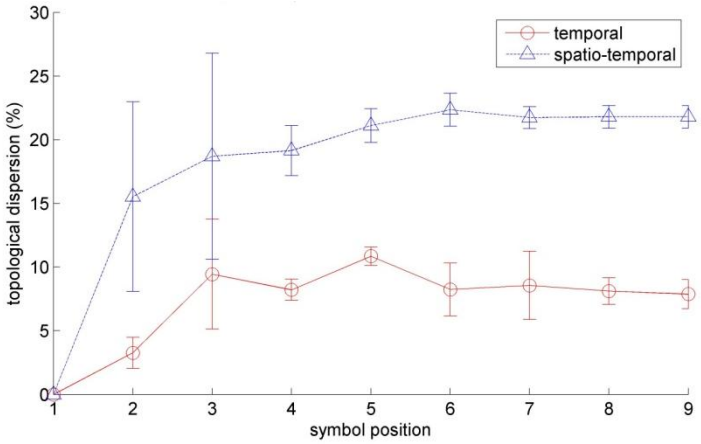

Figure 5: Topological dispersion of symbols on *T*- and *ST*-maps, plotted by their position in input words.

The characteristically distinct ways the two maps recode symbols are reminiscent of the encoding schemes known in the literature as "positional coding" and "Wickelcoding" (Sibley et al. 2008, Davis 2010). It is important to bear in mind, however, that in Hebbian maps symbol encoding is not wired-in in the network's input layer, as customary with connectionist architectures. Rather it is the result of a recoding process based on training data. The number of positional slots, or the length of the left-context affecting symbol recoding is adjusted dynamically on i) the map's memory resources, and ii) the combinatorial complexity of the training input and its frequency distribution.

Differences in the map's inductive bias and recoding scheme have interesting effects on the way input forms are organised and processed through memory structures. *T*-maps are more sensitive to time and can build up stronger expectations over an upcoming symbol in activation. Therefore, they are slightly less accurate than *ST*-maps in perceiving known words (as they trust more their own expectations than actual input stimuli) and considerably less accurate than *ST*-maps in perceiving novel words, for which they built no expectations in the learning phase. When it comes to recalling stored words, however, *T*-maps are more accurate, as they can rely on more accurate positional coding of symbols. On the other hand, *ST*-maps are weaker in capitalising on past events

and thus more tolerant towards unexpected symbols. The implication of this is that they recall novel input sequences more accurately. More importantly for our present concerns, *ST*-maps, unlike *T*-maps, can develop pools of nodes that are specifically sensitive to position-independent *n*-grams. As we shall see in what follows, the notion of position-independent *n*-gram is the closest approximation to the notion of morpheme a Hebbian map can possibly get and has a bearing on the map's ability to recognise morphologically complex novel word forms.

4. Experimental design and materials

To investigate the interplay between prediction and generalisation in the morphological lexicon, we trained instances of a Temporal (*T*) and Spatio-temporal (*ST)* 40×40 T2HSOM on Italian and German text excerpts of about 3,000 word tokens, sampled from two books of child literature: *Pinocchio's Adventures* and the brothers Grimm's *Fairy Tales*. To simulate low-level memory processes for serial order and their impact on morphological organisation, only information about raw forms was provided in training. Such a preliminary step in the process of morphology acquisition is intended to investigate the important but often neglected connection between input word recoding and perception of morphological structure.

Word forms are encoded as strings of capitalised letters preceded by '#' and ended by '$': e.g. '#IST$' for *ist*. Word forms are input to a T2HSOM one letter at a time, with memory of past letters being recoded through re-entrant Hebbian connections that are reset upon presentation of '#'. All letters common to the German and Italian alphabets are written in upper-case. Umlauted characters are written as lower-case digraphs (e.g. '#BRueCKE$' for *Brücke*) and the sharp s '*ß*' as 'ss' (e.g. '#BEIssEN$' for *beißen*). In both cases, pairs of lower-case letters are processed as one symbol. Letters are encoded on the input layer as mutually orthogonal, binary vector codes. Identical letter codes were used for upper-case letters in both German and Italian. 'Five T-maps and five ST-maps were trained on each language for 100 epochs.'. In the five *T*-maps, α= 0.5 and β= 1.0. In the five *ST*-maps, α= 0.087 and β= 1.0. After training, we probed the memory content of the maps on two basic tasks, using both known word forms (i.e. words belonging to the map's training set) and unknown word forms (forming the test set). Both Italian and German test sets contain unseen word forms belonging to word families partially attested in the training set.

4.1. Recoding

The task consists in testing the accuracy of the activation function (1) on both known and unknown word forms. For each symbol *s* shown to the map at time *t*, we test if the map recodes the symbol correctly by activating an appropriate *BMU*(*t*) labelled with *s*. An input word is taken to be recoded accurately if all its letters are recoded accurately. Activation requires faithful memory traces of the currently input symbol, but is also a function of how well the current input symbol is predicted on the basis of past symbols.

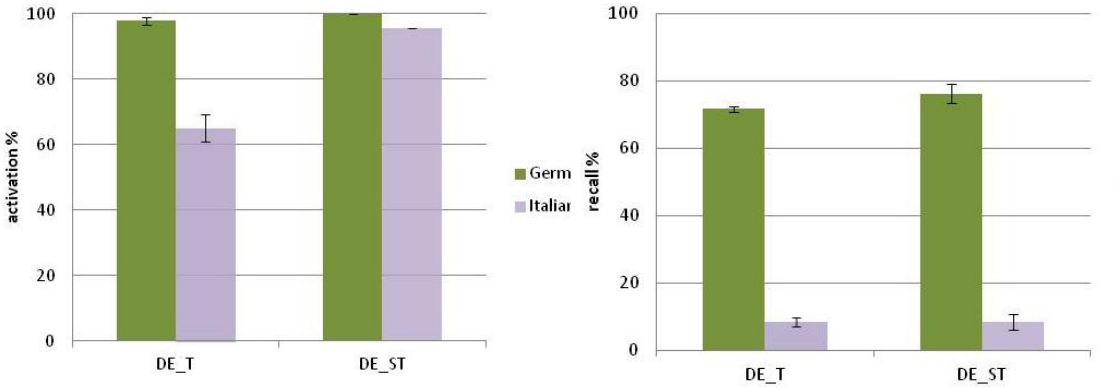

Figure 6. **Left**: Activation scores of German Temporal (DE_T) and Spatio-Temporal (DE_ST) maps, averaged across multiple instances, tested on unknown German words and Italian word forms. **Right**: Recall scores for German *T*-maps and *ST*-maps on unknown German words and Italian word forms.

That more than just storage is involved in recoding is shown by the diagram to the right of Figure 6, providing accuracy scores for both temporal and spatio-temporal German maps, tested on unknown German words and unknown (and unfamiliar) Italian words. Although all Italian letters are present in the German alphabet, 35% of Italian words are wrongly recoded by the German *T*-map (DE_T). This is in striking contrast with the 96% accuracy of German *ST*-maps (DE_ST) on the same task and witnesses the higher sensitivity of *T*-maps to built-in expectations over letter *n*-grams.

4.2. Recall

After Baddeley (1986), we model lexical recall as the task of reinstating a word form soon after a map is exposed to it. The experimental protocol is intended to highlight the dynamic interaction between short-term integration/sustainment of memory traces and long-term storage of lexical information. A *N*-nodes map is first exposed to an input word *w* of length n_w. Its resulting integrated activation pattern $\hat{Y}=\{\hat{y}_1,\ldots,\hat{y}_N\}$, with

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

is input to the same map (n_w*-1)* times. At each time step *t*, the map's *BMU*(*t*) is calculated according to the activation function (1). A word *w* is taken to be recalled accurately if for each *t* ranging from 2 to n_w, the label of *BMU*(*t*) matches the *t*-th letter in *w*. The protocol is thus intended to assess how well the map can output the appropriate sequence of symbols in *w* upon presentation of the whole activation pattern triggered by *w*. Results of recalling words from the training set are shown in Figure 7, grouped by language and map type, and averaged across instances of map type.

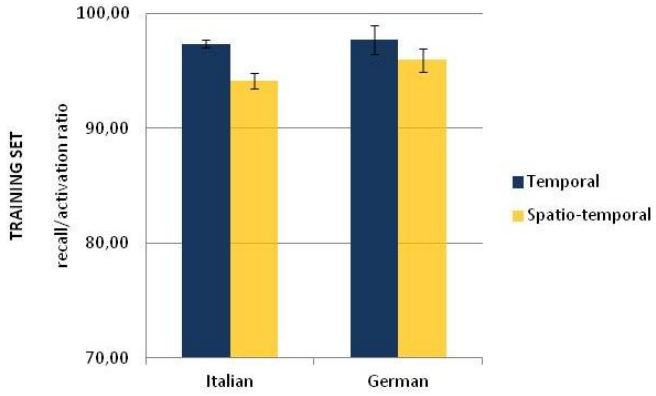

Figure 7: Recall/activation scores on the German and Italian training sets averaged across 5 instances of Temporal and Spatio-Temporal maps.

Note that both temporal and spatio-temporal maps are fairly good at recalling familiar words (training set), with a marginally significant but consistent advantage for temporal maps. This pattern of results is distinctly reversed in Figure 8, plotting the recall/activation ratio on test words, with temporal maps performing consistently worse in both languages. Incidentally, both map types perform considerably worse when tested on recalling unfamiliar unknown words, as is the case of German maps recalling Italian words (see Fig. 6, right).

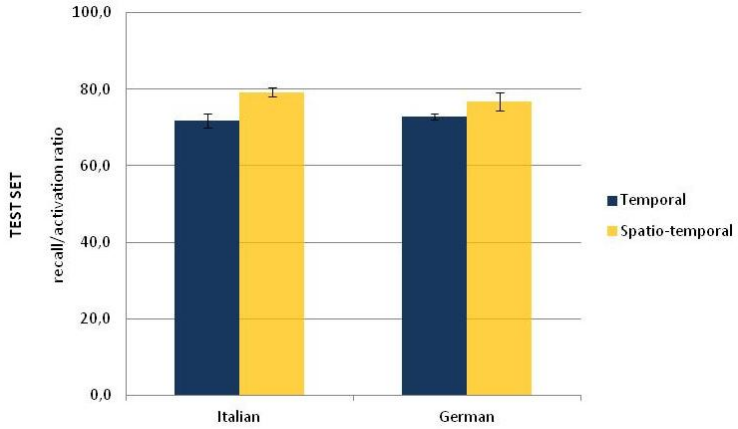

Figure 8: Recall/activation scores on the German and Italian test sets. Scores are averaged across 5 instances of Temporal and Spatio-Temporal maps.

5. Data Analysis

To better understand how generalisation works in Hebbian maps, it is useful to look at Figure 9, where we assume that a (*ST)* map trained on three Italian verb forms (*VEDIAMO* 'we see', *VEDETE* 'you see', and *CREDIAMO* 'we believe') is prompted to recall *CREDETE* 'you believe' afresh. The connection patterns highlighted by grey and red arrows on the trained *ST*-map to the left, are unfolded and vertically arranged in the word graph to the right, to emphasise what is shared and what is not shared by activation patterns.

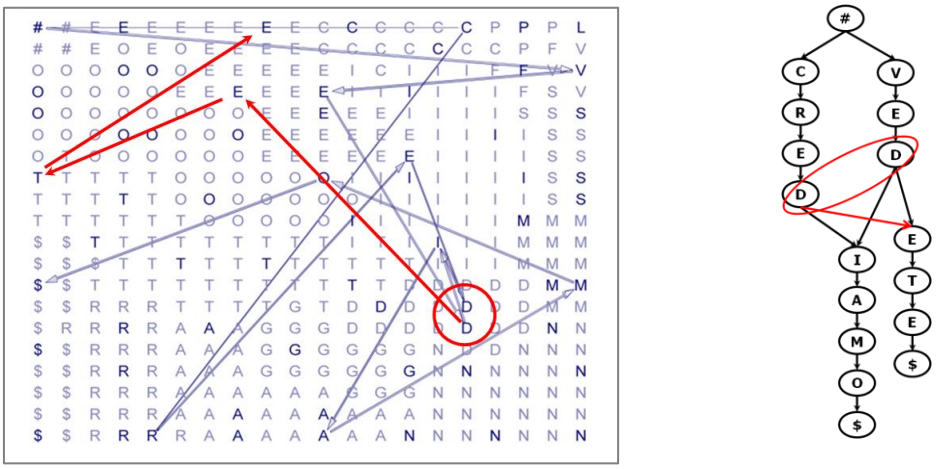

Figure 9: BMU activation chains for *VEDIAMO-VEDETE-CREDIAMO* on a 20×20 map (left) and their word-graph representation (right).

The crucial generalisation step here is represented by a red directed arc in the word graph, and involves the unattested connection between the root *CRED* and the ending –*ETE*. For the map to entertain this connection, it has to be able to i) generalise over the two instances of *D* in *CRED* and *VED*, and ii) align the ensuing ending –*IAMO* in *VEDIAMO* and *CREDIAMO*. The difference in generalisation potential between *T*-maps and *ST*-maps demonstrate that both steps are more likely if symbols are recoded in a context-sensitive but position-independent way, in keeping with the minimal generalisation requirement that rules mapping fully inflected forms are based on the immediate structural context of the change (Albright and Hayes 2003).

Positional encoding appears to be a more effective strategy in lexical recall, suggesting that generalisation and prediction are indeed complementary processing functions, serving different purposes. This is quantitatively summarised in Figure 10 and Figure 11, where we relate the difference in recall accuracy between the two map types to perception of morphological structure, and measures of topological organisation, such as length of *receptive fields*, average per node number of input words and relative number of used-up nodes (*BMU*s).

Following Voegtlin (2002), the receptive field of a map node *n* is defined as the common end of all input strings triggering *n*. For example, if a single node is triggered by '*O*' in the forms *VEDIAMO* and *CREDIAMO* only, its receptive field will be –*EDIAMO*. Accordingly, evidence that *ST*-maps have i) significantly longer receptive fields than *T*-maps (Fig.10, top), ii) more words triggering a single node on average (Fig. 10, centre), and iii) fewer *BMU*s (Fig. 10, bottom), confirms that they are better at finding recurrent substrings in input words. Figure 11 shows how this evidence relates to morphological structure and map's performance. Misalignment (Figure 11, top) tells us how badly activation chains of morphologically-related forms are aligned on shared morphemes (Marzi et al. 2012). High values here indicate that – say – *VEDERE* ('to see') and *CREDERE* ('to believe') activate distinct nodes on their common endings. Lower values indicate a better correlation between activation patterns and shared morphological structure. In turn, this is shown to correlate negatively with accuracy in recalling novel words (Figure 11, bottom). Scores are given for a few verb inflections only.

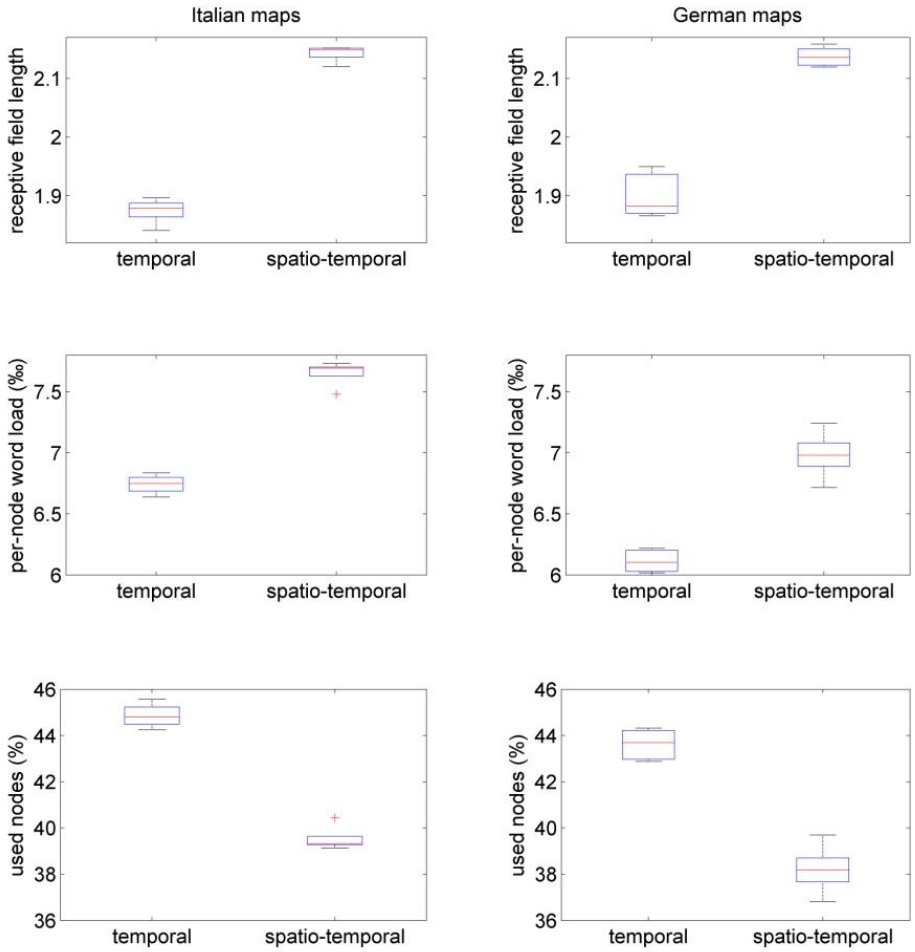

Figure 10: Measures of topological organisation of temporal and spatio-temporal maps on the Italian (left) and German (right) training sets. Scores are averaged across 5 instances of each map type.

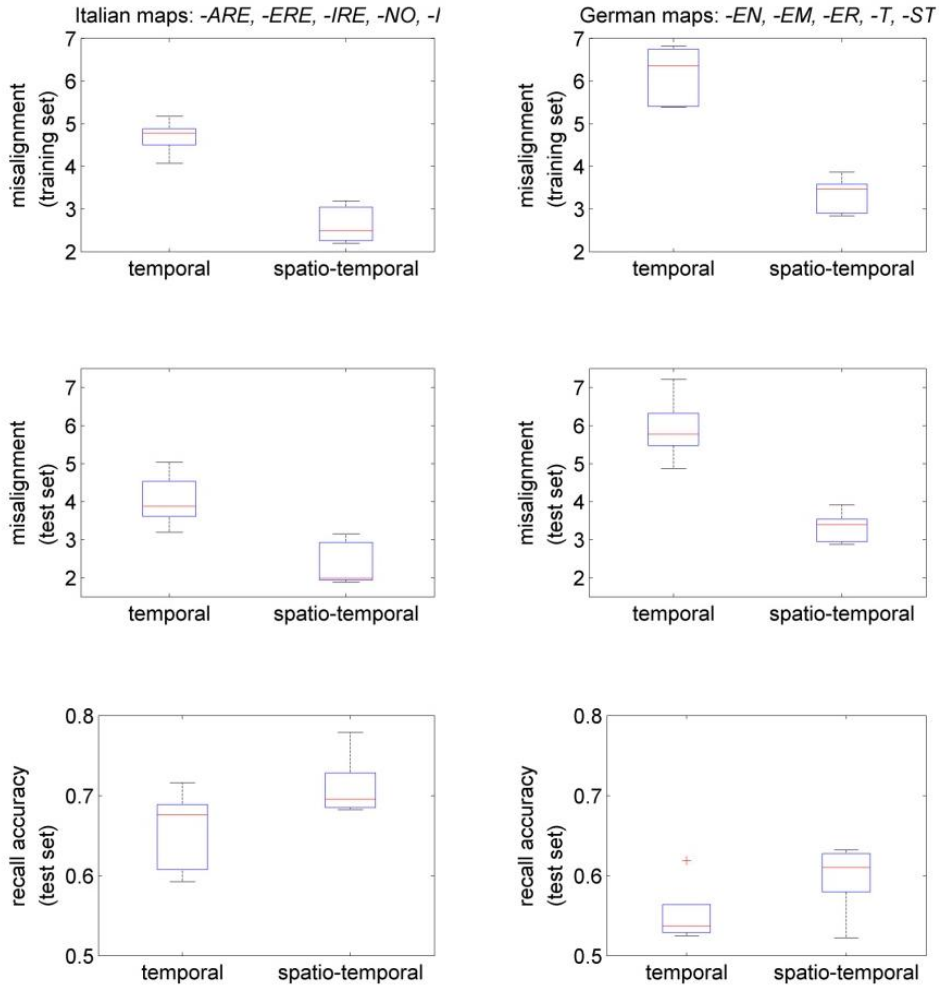

Figure 11: **Top**: misalignment scores across activation patterns triggered by selected inflectional endings on temporal and spatio-temporal maps for both Italian (left) and German (right). **Centre**: misalignment scores for the same set of inflections calculated on the test set. **Bottom**: recall scores of word forms in the test set inflected for the selected endings.

An analysis of the errors made by the two map types in recalling known words indicates that prediction-based errors are more local, involving letter substitution in specific time slots, with no mistakes for letters coming after that slot. Length preservation in the face of local errors is what we would expect for words stored and recalled positionally. This is confirmed by the average per-word percentage of misrecalled symbols in the two maps: about 25% for *T*-maps and more than 37% for *ST*-maps on known Italian words; about 21% for *T*-maps and 34% for *ST*-maps on known German words.

6. Concluding remarks

Prediction affects the way we perceive things and events, through anticipation of upcoming stimuli and integration of missing or noisy information in the input. In Lexical Hebbian maps, prediction is implemented as a process of first-order anticipatory activation of to-be-selected *BMU*s, which presupposes context-driven sensitivity of map nodes to time-bound letters/segments. Thanks to such built-in prediction drive and accurate recoding of time-bound stimuli through training, Hebbian maps show a

remarkable capacity to use past information to process and store the incoming input, offering an interesting model of memory-based word processing and recoding.

In the computational framework offered here, prediction presupposes a bias for past events, under a closed world assumption that what is not currently known to be attested is fairly unlikely, if not impossible. From this perspective, lexical items exhibit minimally redundant patterns which are based on a strictly positional coding of constituent symbols and strong serial connections between them. Morphological generalisation, conversely, seems to require the ability to understand unseen forms based on the discovery of recurrent sub-lexical constituents (morphemes), whose proper coding is context-sensitive but independent of specific positional slots. It is a remarkable aspect of the experimental framework reported here that the two strategies of prediction and generalisation are in fact the outcome of different parameter configurations of a unitary memory model. This has, in our view, a few interesting theoretical implications.

The proposed memory framework radically departs from derivational approaches to morphological competence, by suggesting that principles of lexical organisation may rest on memory self-organisation and recoding, and that rule-like morphological generalisations are the outcome of cautious extension of attested inflections to different word families than those originally attesting them. This move, in our view, blurs not only the traditional linguistic dichotomy between lexicon and rules, but also the related but somewhat more general divide between input/output representations (or knowledge of 'what') and processing principles (or knowledge of 'how'). According to the perspective endorsed in this paper, ways of processing and structural properties of input/output representations are in fact mutually implied, as representations are not given, pre-existing abstract representations but the outcome of an active process of recoding. In turn, processing is memory-driven, with memories of past evidence and already structured information being brought to bear on attentional and combinatorial strategies.

It could be suggested, in line with interactive-activation accounts of word processing (Diependaele et al. 2009), that both strategies for memory organisation (temporal and spatio-temporal) may simultaneously compete in word processing and interact through feedback connections. Temporal maps are better interfaced with the level of lexico-semantic representations, while spatio-temporal maps are more conducive to structured morpho-orthographic representations. Our computational framework allows us to spell out principles of this dynamic interaction in some detail, by putting to extensive empirical test detailed alternative hypotheses. For example, we could test the view that the relative balance between prediction and generalisation is in fact decided dynamically as a function of the stage of acquisition. Item-based learning (Tomasello 2003) may provide an early advantage to children acquiring the lexicon of their own language, as they may find it easier to retrieve and produce a word on the basis of a stronger prediction drive. This can be shown in Figure 12, where we compare the percentage of correctly recalled words by a *T*-map and a *ST*-map through early learning epochs, together with their average length (expressed as a percentage over the average length of all training words). A *T*-map recalls more and longer words at early stages, as item-based storage is relatively local and instantaneous. Finding morphological structure in memorised words, on the other hand, appears to require more time and more evidence for sublexical memory structures to be used in word processing, recoding and retrieval.

Finally, it may well be possible that different languages and morphology types favour either strategy. For example, templatic morphologies and Semitic morphologies in particular may prove to be more conducive to serial, position-based coding of sub-lexical material than more concatenative morphologies are, suggesting that a bias for either strategy could develop as the result of learning. Algorithms for morphology acquisition 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.

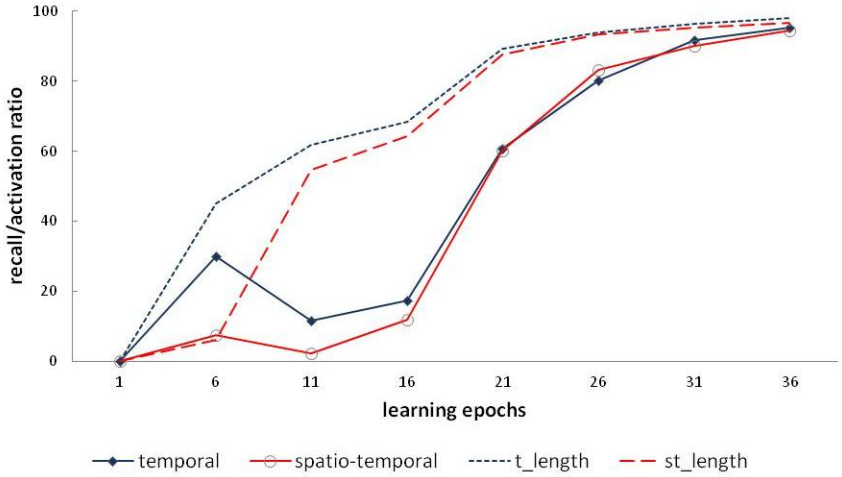

Figure 12: Recall/activation scores and average length of recalled words by early learning epochs for *T*-maps and *ST*-maps.

References

Albright, A., Hayes, B. (2003), Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90, 119–161.

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

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

Baddeley, A.D., and Gathercole, S.E. (1992), Learning to read: The role of the phonological loop. In J. Alegria, D. Holender, J.J. de Morais and M. Radeau (eds.), *Analytic approaches to human cognition*, Amsterdam, Elsevier, 153-168.

Berko, J. (1958), The child's learning of English morphology. *Word* 14, 150-77.

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

Booij, Geert (2010), Construction Morphology. Oxford, Oxford University Press.

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.

Bybee, J.L., Pardo E. (1981), Morphological and lexical conditioning of rules: Experimental evidence from Spanish. *Linguistics* 19, 937-68.

Bybee, J.L., Slobin D.I. (1982), Rules and Schemas in the development and use of the English Past Tense. *Language,* 58, 265-289.

Bybee, J.L., C. Moder 1983 "Morphological Classes as Natural Categories" *Language* 9, 251-270.

DeLong, K.A., Urbach, T.P., Kutas, M., 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8, 1117-1121.

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

Corbett, G., and Fraser, N. (1993), Network Morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics*, 29, 113-142.

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

Davis, C. J. (2010), The Spatial Coding Model of Visual Word Identification. *Psychological Review*, 117, 3, 713-758.

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

Diependaele, K., Sandra, D., & Grainger, J. (2009). Semantic transparency and masked morphological priming: The case of prefixed words. *Memory & Cognition*, 37(6), 895-908.

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

Ferro M., Pezzulo G., Pirrelli V. (2010), Morphology, Memory and the Mental Lexicon. In *Lingue e Linguaggio*, V. Pirrelli (ed.) *Interdisciplinary aspects to understanding word processing and storage*, Bologna, Il Mulino, 199-238.

Ferro, M., Marzi, C., Pirrelli, V. (2011), A self-organizing model of word storage and processing: implications for morphology learning, in *Lingue e Linguaggio*, X, 2, Bologna, Il Mulino, 209-226.

Ford, M., Marslen-Wilson, W., 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.

Gathercole, S.E., Pickering, S.J. (2001), Working memory deficits in children with special educational needs. *British Journal of Special Education*, 28, 89-97.

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

Maratsos, M. (2000), More overregularizations after all. *Journal of Child Language*, 28, 32-54.

Marzi, C., Ferro, M., Caudai, C., Pirrelli, V. (2012), Evaluating Hebbian Self-Organizing Memories for Lexical Representation and Access. In *Proceedings* of LREC 2012, Istanbul, Turkey.

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

McClelland, J., Patterson, K. (2002), Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Science*, 6, 465-472.

Pickering, M.J., Garrod, S. (2007), Do people use language production to make predictions during

comprehension? *Trends in Cognitive Sciences*, 11, 105-110.

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

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

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

Pirrelli, V., Ferro, M., Calderone, B., (2011), 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, 135-157.

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

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

Rumelhart, D. E., McClelland, J. L. (1986), On learning of past tenses of English verbs. In J.L. McClelland and D.E. Rumelhart (eds.), *Parallel distributed processing*, Cambridge, MA, MIT Press, 2, 216-271.

Seidenberg, M.S., 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., Elman, J.L. (2008), Large-scale modeling of wordform learning and representation, *Cognitive Science*, 32, 741-754.

Smolensky, P. (1988), On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11, 1-74.

Staub, A., Clifton, C., Jr (2006), Syntactic prediction in language comprehension: evidence from either. . .or. *Journal of Experimental Psychology, Learning, Memory and Cognition*, 32, 425–436.

Stemberger, J.P., Middleton C.S. (2003), Vowel dominance and morphological processing, *Language and Cognitive Processes*, 18(4), 369-404.

Tabak, W., Schreuder, R., Baayen, R.H. (2005), Lexical statistics and lexical processing: semantic density, information complexity, sex and irregularity in Dutch. In M. Reis and S. Kepser (eds.), *Linguistic Evidence*, Berlin, Mouton de Gruyter, 529-555.

Tomasello, M. (2003), Constructing a Language: A Usage-based Theory of Language Acquisition. Cambridge, MA, Harvard University Press.

Wilson, M., Knoblich, G. (2005), The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131, 460-473.