Effects of frequency and regularity in an integrative model of word storage and processing

Claudia Marzi, Marcello Ferro, Franco Alberto Cardillo, Vito Pirrelli

Institute for Computational Linguistics, National Research Council, Via G. Moruzzi 1, 56126 Pisa, Italy <claudia.marzi@ilc.cnr.it>, <marcello.ferro@ilc.cnr.it>, <francoalberto.cardillo@ilc.cnr.it>, <vito.pirrelli@ilc.cnr.it>

Considerable evidence has accrued on the role of paradigms as both theoretical and cognitive structures regimenting the way words are processed and acquired. The evidence supports a view of the lexicon as an emergent integrative system, where word forms are concurrently and competitively stored as repeatedly successful processing patterns, and on-line processing crucially depends on the internal organisation of stored patterns.

In spite of converging evidence in this direction, little efforts have been put so far into providing detailed, algorithmic models of the interaction between lexical token frequency, paradigm frequency, and paradigm regularity in word processing and acquisition. Here we propose a neuro-computational account of the frequency/regularity interaction, and discuss some of its theoretical implications by analysing experimental results in the computational framework of Temporal Self-Organising Maps. Detailed quantitative analysis shows that the model provides a unitary explanatory framework bringing together insights from neighbour family effects on word recognition and production, evidence from family size effects in serial lexical access and paradigm-based dynamics in lexical acquisition.

KEYWORDS: Lexical access, word recall, serial processing, parallel activation, inflectional paradigms, mental lexicon.

1. Introduction

In spite of converging evidence on the role of morphological families and paradigmatic relations in the developmental course of lexical acquisition and processing, there have been no attempts to simulate the interdependency between simple mechanisms of lexical activation/competition and effects of lexical token frequency, paradigm frequency, and paradigm regularity in word processing and acquisition. One of the fundamental limitations of most existing computational models of word recognition and production (McClelland & Elman 1986; Norris, McQueen & Cutler 1995; Levelt et al. 1999; Gaskell & Marslen-Wilson 2002; Chen & Mirman 2012; among others) is that they either focus on processing issues, by analysing how input patterns can be mapped onto existing stored exemplars during process-

ing, or focus on storage, by entertaining different hypotheses concerning stored representations and how they are affected by differences in input data distribution. We appear to be missing more 'Integrative' (neuro)computational models of the mental lexicon (Marzi & Pirrelli 2015), where (i) structures that are repeatedly activated in processing an input word are the same units responsible for its stored representation, and (ii) they are made develop dynamically as the result of learning. In our view, truly integrative models would lead to a better understanding of the dynamic interaction between processing and storage, and make room for a careful analysis of the empirical consequences of such a mutual implication on a sizeable amount of realistic lexical data. This is the goal of the present study. By running two simulations of the acquisition of verb paradigms in German and Italian through temporal self-organising artificial neural networks (or Temporal Self-Organising Maps, TSOMs), we investigate the time-bound dynamics of co-activation and competition in the acquisition of families of inflected data. In addition, we examine whether the same basic principles can correctly predict the direction of lexical neighbour effects on lexical access and production of the same data. Quantitative and qualitative analyses of our experimental results show that a unitary account of paradigm-based lexical acquisition and processing effects of neighbour families is possible, and that both acquisition and processing effects are amenable to independently motivated computational principles of Hebbian learning.

Firstly, we sketch the theoretical background (Section 2) of the present work, and the computational architecture (Section 3) adopted for our simulations. Experimental results are then illustrated and analysed with linear mixed effects models (Section 4). A general discussion (Section 5) follows, summarising our results in the framework of an integrative model for task-based memory and processing strategies.

2. Theoretical background

Families of inflectionally-related words (be they word paradigms such as *walk*, *walks*, *walking*, *walked*, or classes of identically inflecting forms such as *walking*, *playing*, *reading*, etc.), or derivationally-related words (e.g. *form*, *formation*, *formal*, *formalize*, *formalization*, etc.), have received increasing attention over the last 25 years, both in the theoretical literature on morphological competence, and in the psycho-linguistic debate on the organising principles of the mental lexicon. In particular, considerable emphasis has been laid on the role

of paradigmatic relations as principles of non-linear organisation of word forms in the speaker's mental lexicon, facilitating their access, retention and use (Baayen et al. 1997; Orsolini & Marslen-Wilson 1997; Bybee & Slobin 1982; Bybee & Moder 1983; among others).

In line with so-called 'Words and Paradigms' approaches to morphological competence (Blevins 2006 among others), mastering the inflectional system of a language amounts to acquiring an increasing number of paradigmatic constraints on how paradigm cells should be filled in (see Ackerman et al. 2009; Finkel & Stump 2007; Pirrelli & Battista 2000; Matthews 1991; among others). The view is supported by growing evidence of the role of morphological paradigms in the developmental course of word acquisition. Children are shown to be sensitive to sub-regularities holding among paradigm cells (see, for a comprehensive picture, Bittner et al. 2003; on Italian, Orsolini et al. 1998; Colombo et al. 2004; on Polish, Dabrowska 2004, 2005). Contrary to both rule-based (e.g. Pinker & Ullman 2002; Albright 2002) and most connectionist simulations of word acquisition (see Rumelhart & McClelland 1986; MacWhinney & Leinbach 1991; Plunkett & Juola 1999; among others), little evidence supports the assumption that one underlying base form can be used to produce, on line, all inflected forms of a paradigm. Rather, the relational structure of all forms of a paradigm appears to enforce global, distributed constraints on both word acquisition and processing, with redundant relations and multiple 'bases' playing a fundamental role in lexical competence (Burzio 2004). According to this view, the mental lexicon is an emergent integrative system, where words are concurrently, redundantly and competitively stored (Alegre & Gordon 1999; Baayen 2007; Bybee 1995). No categorical distinction is made between regular and irregular inflected forms, nor between uniquely stored bases and non-base forms, seemingly derived by speakers on demand (see Baayen 2007; Marzi 2014 for a recent overview). An interesting computational consequence of this view is that storage and processing are mutually implied. First, to capture the fact that words encountered frequently exhibit different lexical properties from words encountered less frequently, any model of lexical access must assume that accessing a word in some way affects the access representation of that word (e.g. Forster 1976; Marslen-Wilson 1993; Sandra 1994). Accordingly, entries for high-frequency words are assumed to exhibit higher levels of resting activation and are typically associated with entrenched, whole-word memory representations. In contrast, low-frequency words are associated with weaker and more distributed lexical representations, accounting for their complex morphological structure

(Schreuder & Baayen 1997; Baayen & Schreuder 1999, 2000). Even though word forms are memorised in the lexicon irrespective of their degree of morphological complexity, not all of them are memorised equally: their representations in fact reflect the way words are processed, with levels of entrenchment and levels of resting activation being a function of the probabilistic support words receive from repeatedly successful processing steps.

Secondly, processing is in turn based on existing memory structures. This is the tenet of so-called 'memory-based' models of language processing (e.g. Daelemans & van den Bosch 2005), according to which analogy-based re-use of stored examples is more suited for language processing than the application of rules extracted from those examples. The approach has received support from recent advances in understanding the neuro-anatomical areas supporting memory (Wilson 2001; D'Esposito 2007; Ma et al. 2014), showing that working memory consists in the transient activation of long-term memory structures, controlled and maintained by the integration of auditory-motor circuits in the perisylvian network (Catani et al. 2005; Shalom & Poeppel 2008). This can explain speed-up effects in processing high-frequency words. Words that are seen more often will activate the same memory circuits over again, whose strength is increased as a function of repeated usage. As a result, some circuits gradually specialise to respond to some input words only, increasing their speed and processing efficiency. Likewise, words in large, densely interconnected word families are reacted to by speakers more quickly in lexical decision tasks (Baayen et al. 1997; Ford et al. 2003; Lüdeling & de Jong 2002), with words belonging to large, highly entropic, inflectional paradigms being accessed faster and more accurately than words in smaller paradigms (Moscoso 2007; Moscoso et al. 2004).

The cognitive literature on similarity-based principles of word association has greatly contributed to understanding effects of family size and frequency of neighbouring words on a variety of word processing tasks: non-word repetition (Vitevitch et al. 1997; Vitevitch & Luce 1998), recall from verbal short-term memory (Gathercole et al. 1997), phoneme identification (Pitt & McQueen 1998) and word recognition (Luce 1986; Luce & Pisoni 1998). Beyond specific differences depending on the nature of the input stimuli (e.g. acoustic vs. visual) and the processing requirements of the task (e.g. word recognition vs. word production), an interesting general pattern of reversal emerges: neighbours have facilitative effects on spoken word production and inhibitory effects in spoken word recognition. Furthermore, the frequency distribution of neighbours plays an important role in

determining whether competition/co-activation effects are facilitative or inhibitory: high-frequency neighbours tend to exert an inhibitory effect on some processing tasks, while low-frequency neighbours facilitate execution of the same tasks.

3. TSOMs

Temporal Self-Organising Maps (or TSOMs) are a variant of classical Kohonen's SOMs (Kohonen 2001) specifically designed for processing and storing time-bound series of symbols. A TSOM consists of a grid of fully interconnected processing nodes that concurrently activate in response to input symbols shown in temporal contexts (Koutnik 2007; Ferro et al. 2010a; Pirrelli et al. 2011; Marzi et al. 2012; Marzi et al. 2014). Map nodes mimic neural receptors that are trained to get increasingly sensitive to specific time-bound input signals.

In TSOMs, learning consists in the topological (pattern matching) and temporal (pattern synchronisation) co-organisation of connection weights on multiple levels of connectivity (Figure 1). Hebbian rules are applied at all levels, so that nodes highly responsive to a stimulus (e.g. a given input symbol in a given context) will be even more responsive to that stimulus as training goes on. Conversely, nodes weakly responsive to a stimulus, will be even less responsive to that stimulus. After training, each node in a TSOM can be labelled with the input symbol the node responds most strongly to.

3.1. The architecture

A TSOM is a grid of processing nodes with multiple levels of weighted connections, propagating information with different time delays (Figure 1). Input connections are used to get synchronous information from an input layer, where individual stimuli are sampled at one-time tick. The amount of information conveyed by each connection is a direct function of its weight, ranging in the [0; 1] interval. Temporal connections, on the other hand, simulate neuron synapses with one-tick delay propagation. Their weights determine the amount of influence that activation of one node at time t has on the activation of nodes at time $t + 1$. Temporal connections thus convey the probabilistic expectation that any map node is about to be activated, given the current activation state of the map.

When a symbol is presented on the input layer, all nodes are fired synchronously. Their activation is the result of a weighted summation of signals on both input and temporal connections. For each

node on the map, its processing response to an input stimulus at time is given by:

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

where $y_{S,i}(t)$ is the amount of activation to the i-th node at time t flowing through input connections, and $y_{T,i}(t)$ is the temporal activation of the i-th node at time t triggered by the state of activation of all nodes at time $t - 1$. In the equation, α and $(1 - \alpha)$ weigh up the respective contribution of input connections (S) and temporal connections (T) to activation of node i. More intuitively, equation (1) integrates the state of map's activation caused by the current input symbol with the amount of expectation raised by experiencing the immediately preceding symbol.

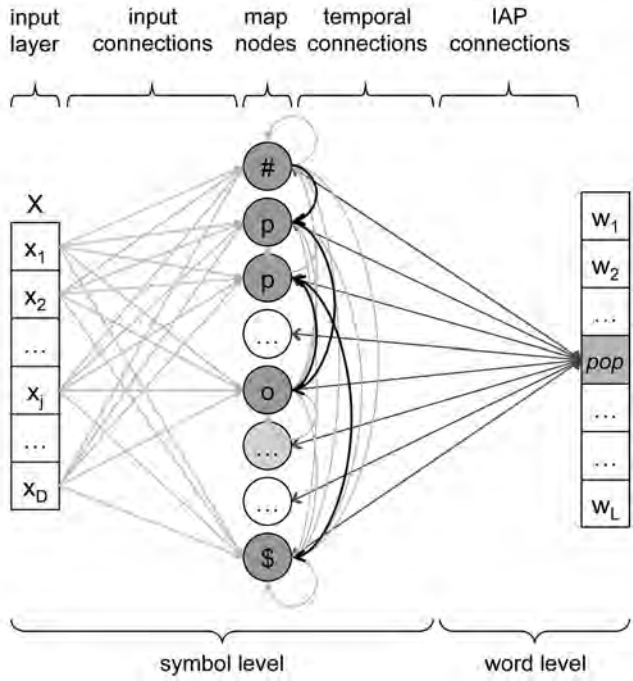

Figure 1. Outline architecture of a TSOM. Map nodes show the Integrated Activation Pattern for the input string '#pop$'. For simplicity, only BMU (Best Matching Unit) nodes are labelled and connected through edges/arcs. Shades of grey depict levels of node activation. Forward temporal connections between BMUs are highlighted as black arcs.

At each ensuing time tick, a new symbol is shown on the input layer. The newly generated pattern of node activation is recurrently integrated with the previous activation state of the map. When a time series of input symbols is terminated, the resulting Integrated Activation Pattern (or IAP) of nodes represents the processing response of the map to the whole input series. Figure 1 illustrates an IAP for the input sequence '#pop$', where '#' and '$' mark, respectively, the start and the end of the sequence. Shades of grey pictorially represent levels of node activation, corresponding to values $\hat{Y} = \{\hat{y}_1, ..., \hat{y}_N\}$, where $\hat{y}_i$ is the maximum level of activation reached by node responding to '#pop$', namely:

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

where i ranges over N map nodes, and t over symbol positions in the input string.

Connection weights are not wired-in, but trained after presentation of each new input signal. Following synchronous activation of the map by an input signal, the most highly activated node, or Best Matching Unit (BMU), is trained in two steps. First, weights on all input connections to BMU are adjusted for them to be closer to the current input signal. Likewise, all temporal connections to BMU are made more correlated with the overall activation pattern of the map at time $t - 1$. In particular, the weight on the connection between BMU at $t - 1$ and the current BMU is increased (potentiation, Figure 2, left); the weights on the connections between all other nodes and the current BMU are decreased (inhibition, Figure 2, left). Secondly, weight adjustment spreads radially to neighbour nodes with a Gaussian function centred on the current BMU. Radial propagation prompts information sharing and training dependence between topologically adjacent nodes, which are thus trained to respond alike to similar input stimuli (Pirrelli et al. 2015).

The two training steps ensure selective specialisation of map nodes. Nodes get gradually more sensitive to specific time-bound instantiations of input symbols. For example, given an input bigram ab, the connection weight between BMU for a at time $t - 1$ and BMU for b at time t increases when a precedes b. The same connection weight decreases when b is preceded by a symbol other than a (see Figure 2, left). Due to this dynamic, if ab is a high-frequency input bigram, the map will develop a specialised node for b in ab, i.e. a node that is selectively activated each time the BMU for a is activated at the immediately preceding time tick (see Figure 2, right). Conversely,

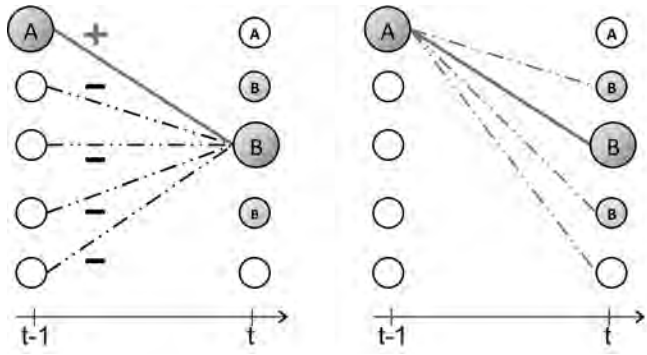

Figure 2. Left: Graphical representation of rules for temporal Hebbian learning: the '+' edge stands for a potentiated connection, '–' edges for inhibited connections. Right: Range of forward one-tick-delay connections leaving node 'A' at time $t - 1$. The solid edge represents the strongest connection, and dashed edges represent weaker connections. In both graphs, nodes with larger labels represent BMUs at consecutive time ticks. Shades of grey indicate levels of node activation.

lower frequency bigrams tend to activate less specialised BMUs, or 'blended' BMUs (Marzi & Pirrelli 2015), meeting the requirements of different time-bound instances of the same symbol.

Selective specialisation of map nodes propagates through time. Given the trigram *abc* being repeatedly input to a TSOM, the map will first develop a specialised BMU responding to *b* following *a*. In turn, the BMU will strengthen a temporal connection to another dedicated BMU responding to *c* following *b*. In general, any sequence of symbols can be associated with a specialised integrated pattern of BMUs (IAP) depending on its own relative input frequency, the map's plasticity and availability of map nodes (Pirrelli et al. 2011; Marzi & Pirrelli 2015).

Turning back to the IAP of Figure 1, the top-most activated units are, by definition, the BMUs responding to '*#*' in first position, *p* in second position, *o* in third position, *p* in fourth position and '*$*' in fifth position in the input string. Note that, due to selective specialisation and radial propagation, two distinct, topologically close BMUs are recruited to respond to the same symbol *p* in different contexts. Finally, the IAP is synchronised with a localist word-level node through a layer of IAP connections, keeping long-term memory of the activation pattern in their weights.

3.2. Using TSOMs as lexical maps

When a TSOM is trained on a set of word forms, weights on all connectivity layers are adjusted in an experience-dependent way, as a function of the frequency distribution and the amount of formal redundancy in the training data. After an initial period of random variability, where nodes activate chaotically and inter-node connections are distributed uniformly, a map gradually develops increasingly specialised IAPs for words in the training set. Thanks to the interplay between selective specialisation and radial propagation, a TSOM apportions its processing resources for them to respond more strongly to more frequent input words, and less strongly to less frequent words. This is the result of strengthening repeatedly used connections, which are specialised for processing highly expected input signals. Weaker resources are kept for less frequent but typical words.

Due to this bias, high-frequency words tend to be associated with entrenched IAPs, whose BMUs are strongly connected with one another, and weakly connected with any other nodes. Specialised inter-node connectivity makes BMUs more salient and less confusable, as they receive stronger support through temporal connections than any other node. The same is true of BMUs responding to formally atypical words in the lexicon, i.e. words surrounded by few or no lexical neighbours. Since they are fairly isolated, atypical words are likely to activate fairly specialised BMUs.

Low-frequency typical words, on the other hand, are associated with 'blended' BMUs, which are densely and weakly connected with many other nodes, to meet the input requirements of more words. Because of poorly selective connectivity, their levels of activation are more evenly spread through their IAP, thus suffering the competition of other non-BMU nodes in the same IAP, and the parallel activation of other IAPs associated with similar input words.

When TSOMs are trained on highly redundant input data such as verb paradigms, specialisation and blending may interact. Figure 3 gives a graph-like representation of the possible temporal connectivity of BMUs responding to some verb forms of German *glauben* 'believe'. A pool of shared BMUs is associated with the common verb stem *glaub-*, their temporal connections being strengthened each time any form of glauben is input to the map. In addition, the specialised *glaub-* sub-pattern is connected with many inflectional endings through a blended range of one-to-many forward connections (see Figure 3). Upon activation of the *b* node at time *t*, the map propagates the activation through its forward temporal connections and

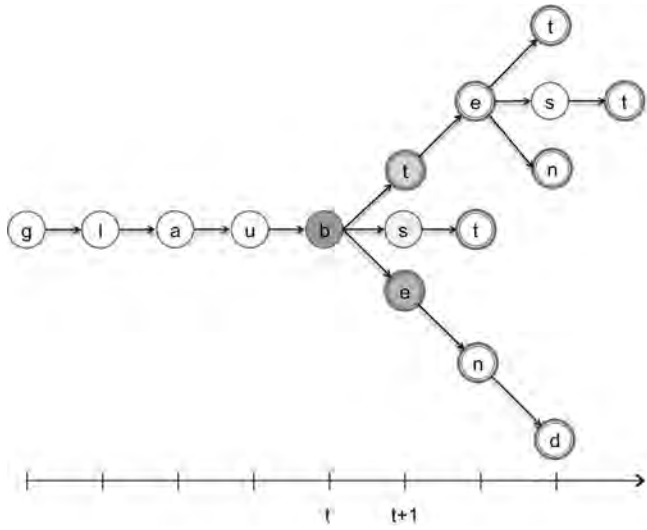

Figure 3. Graph-like representation of the temporal connectivity of BMUs associated with some inflected forms of *glauben*. BMUs responding to word final symbols are double-circled. Shades of grey depict levels of node activation at time *t*. The BMU at time *t* propagates its expectations to prospective BMUs (time *t* + 1), whose varying levels of pre-activation reflect the probabilistic support they receive from the current BMU. Only relevant connections are shown.

prospectively co-activates up-coming nodes. Ultimately, activation of the blended IAP of any form of *glauben* co-activates IAPs for the other members of the same paradigm, which will in turn compete for selection. Hence, activation of a sub-pattern shared by members of the same paradigm prompts paradigm-sensitive co-activation and competition of blended IAPs.

Owing to this dynamic, IAPs are not only short-term processing responses of the map to input words. The long-term knowledge sitting in BMUs' connections makes IAPs also routinized memory traces of the same processing responses. Given an IAP and the temporal connections between its BMUs, a TSOM can use this knowledge to predict, for any currently activated BMU in the IAP, the most likely upcoming BMU. This makes it possible to test the behaviour of a TSOM on two classical lexical tasks: serial word access and word recall. The two processes are simulated as described in the following sections.

3.2.1. Serial word access

In serial word access, we simulate how the map can predict an incrementally presented input word. After training, each word in the training set is progressively presented to a TSOM by showing one symbol at a time on the input layer. Upon each symbol presentation, the TSOM is prompted to complete the current input string, by anticipating its possible continuation. A TSOM can predict a progressively presented input word by propagating activation of the current BMU through its forward temporal connections, and outputting the label associated with the most strongly (pre)activated node:

$$(3) \qquad BMU\,(t+1) = \underset{i=1,\ldots,N}{\operatorname{argmax}} \{m_{i,h}\} \qquad h = BMU\,(t)$$

where $m_{i,h}$ is the weight value on the forward temporal connection from the node h to the node i. Each correctly predicted symbol in the input word is assigned the prediction score of the preceding symbol incremented by 1. Otherwise, the symbol receives a 0-point score.

In the experimental results reported in Section 4.3, we averaged per-symbol prediction scores across the input word's length, to reflect how 'wordlike' the input word is, i.e. how typical with respect to other words in the lexicon (Bailey & Hahn 2001). This is a measure of global, lexical familiarity, and depends on how many neighbours the input word has in the lexicon, irrespectively of its own level of memory entrenchment.

3.2.2. Word recall

Given a word's IAP, we can use it as an input activation pattern to test how well the map can retrieve the word from its pattern. We simulate this by letting the map go through a word IAP, and iteratively output, at each time tick, the label of the current BMU. Since an IAP is a static pattern of synchronously activated nodes (equation 2, Section 3.1), the task tests how accurately levels of node activation in the IAP encode information about the timing of the input symbols that make up the word. The process of recall consists in: (i) prompting the map with the start-of-word symbol ('#'), (ii) integrating the IAP with the current temporal expectations and calculating the BMU, (iii) repeating step (ii) over again until the end-of-word symbol ('$') is reached. A word is recalled correctly if all its symbols are recalled correctly in the appropriate order.

More formally, the map iteratively processes the IAP as an input activation pattern according to:

$$y_i(t) = \alpha \cdot \hat{y}_i + (1 - \alpha) \cdot y_{T,i}(t) \quad (4)$$

where, at each time t, the most highly activated i node is the result of integrating information in the current IAP ($\hat{y}_i$ in Equation 4) with the dynamically updated expectations of the map ($y_{T,i}(t)$).

Some IAPs are more confusable than others. The recall of a word from its IAP can be more or less easy depending on the degree of co-activation of other non-target IAPs whose BMUs are highly activated in the target IAP. For example, if two input strings present some symbols in common (e.g. *write* and *written*, *macht* and *gemacht*), they will tend to activate largely overlapping patterns of nodes. To counteract the potential interference of spurious BMUs during recall, a map can filter out nodes whose level of (co-)activation does not reach a set threshold. The stronger the competition of potential intruders, the higher the threshold needed to filter them out. The amount of filtering (or threshold level) required can thus tell us how difficult it is for the map to recall the target word.

4. Experimental evidence

4.1. Data and design

Fifty German and fifty Italian verb sub-paradigms were selected among the most highly ranked paradigms by cumulative frequency in a reference corpus (CELEX Lexical database for German, Baayen et al. 1995; Paisà Corpus for Italian, Lyding et al. 2014), to study the dynamics of word and paradigm acquisition in German and Italian verb inflection.

For each paradigm, an identical set of 15 cells was used for training, for an overall number of 750 inflected forms for each language. Each data set was administered to the map for 100 epochs under two different training regimes: a uniform distribution (UD: 5 tokens per word), and a function of real word frequency distributions in the reference corpus (skewed distribution or SD: with token frequencies in the range of 1 to 1001). For each training regime, we ran 5 TSOM instances. More details on dataset distribution and composition are given in Appendix.

By varying frequency distributions and comparing the effects of inflectional complexity of training data on lexical access, word recall and word acquisition, we wanted to gain some insights into the interplay between morphological regularity and word frequency. After training, we monitored the behaviour of the resulting TSOMs (namely

UD Italian, SD Italian, UD German and SD German, over 5 different instances to then average our results) by inspecting the time of acquisition of words and paradigms. For this purpose, we define the time of acquisition of a single word as the training epoch whence a TSOM can accurately recall the word in question from its IAP. Likewise, for each paradigm, its time of acquisition by a map is the mean acquisition epoch of all forms belonging to the paradigm.

4.2. Word and paradigm acquisition

As a general trend, TSOMs acquire word forms by token frequency, with higher-frequency words being quickly memorised and successfully recalled at earlier learning epochs, as shown in Figure 4, where token frequency is averaged over words that are correctly recalled at each learning epoch.[1] This is not surprising, given the dynamic of selective specialisation illustrated in Section 3. A highly frequent input string tends to repeatedly activate the same pattern of nodes, strengthening the connections between sequentially activated BMUs, and establishing a dedicated, highly responsive IAP.

When it comes to the actual timing of paradigm acquisition, however, things get considerably more complex, with the notion of morphological regularity interacting non-trivially with token frequency distributions. In both German and Italian, the vast majority of paradigms are acquired significantly earlier[2] ($p < .005$) in a UD regime than in an SD regime (Figure 5). All in all, paradigm

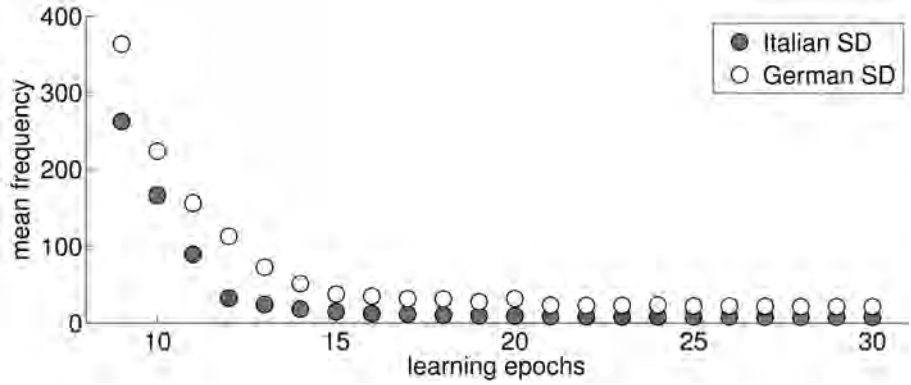

Figure 4. A scatter plot of the mean token frequency of correctly recalled words for the first 30 learning epochs of two TSOMs trained on Italian (black circles) and German data (white circles) in a skewed regime (SD).

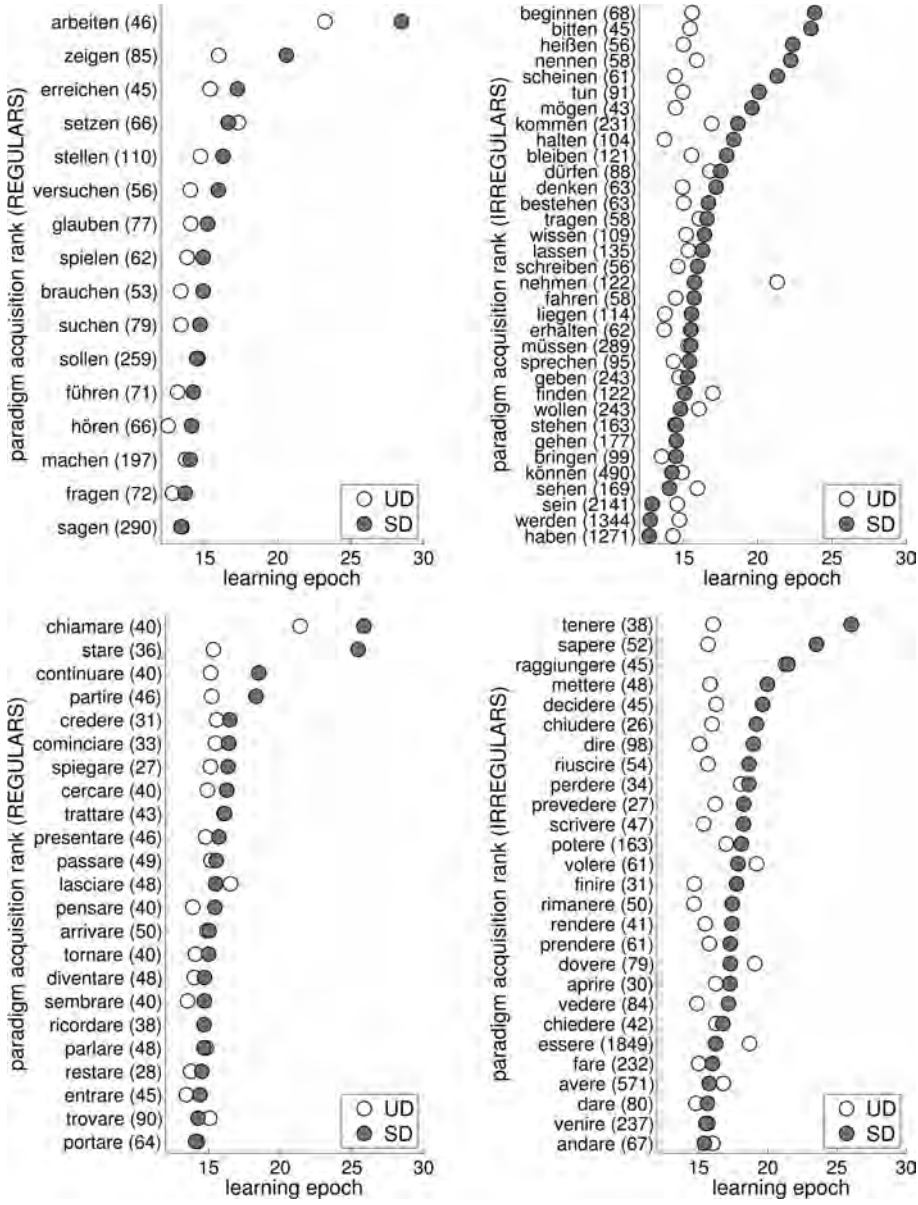

Figure 5. Time course of regular (left) and irregular (right) paradigms ranked top-down by decreasing learning epoch; results are shown for both skewed distributions of training data (SD: grey circles) and uniform distributions (UD: white circles), for German (top) and Italian (bottom). Values are averaged across 5 map instances for each type. Paradigm cumulative frequencies are given in brackets.

acquisition does not appear to be solely influenced by token frequency effects. What particularly matters is how token frequencies are distributed both within each paradigm and across paradigms, with relative frequency distributions causing more global effects on ease and speed of acquisition. Only a handful of high-frequency irregular German paradigms make exception to this trend: the paradigms of *sein* 'be', *werden* 'become' and *haben* 'have' (among others) are found to be learned earlier when they are input to the map according to corpus-based distributions (SD regime). This seems to further suggest that frequency and regularity do not interact linearly, and that the effects of frequency on paradigm acquisition are crucially affected by the degree of morphological regularity and formal redundancy that words exhibit at the level of their inflectional paradigm.[3]

In the German data, regular paradigms are less sensitive to token frequency effects and to differences in token frequency distributions than irregular paradigms are, as witnessed by the strong correlation[4] ($r = .92$, $p < .00001$) between the time courses of acquisition of regular paradigms in UD and SD regimes (Figure 5, top, left panel). That token frequency affects the acquisition of words in regular paradigms to a lesser extent than the acquisition of irregular ones can be explained by observing that a TSOM takes advantage of the cumulative frequency of stems across the whole paradigm. Words in regular paradigms, in fact, exhibit a statistically significant correlation between stem cumulative frequency and time of acquisition ($r = -.36$, $p < .00001$).

This is not just a memory effect based on cumulative stem frequencies. In both languages, regular paradigms tend to be acquired earlier (in terms of significantly earlier learning epochs, $p < .01$) in the UD training regime. Besides, in both training regimes, regular paradigms are acquired more quickly than irregular paradigms are, as they appear to be associated with significantly shorter learning spans ($p < .005$), i.e. lower number of epochs between the acquisition time of the first and the last member of a paradigm. It looks like uniformly trained maps are able to organise stored words in a more deeply interconnected network of associative relations, where nodes share information through distributed patterns of poorly specialised connections. Information sharing favours co-activation (i.e. spreading of activation to other neighbouring/similar forms), perception of internal structure and, ultimately, generalisation across word families (Marzi et al. 2014).

4.3. Frequency by regularity interaction

Many of the effects reported in the previous section are the dynamic result of two interacting dimensions of memory self-organisation in TSOMs: (i) the linear dimension characterises serial processing, and controls the level of predictability and entrenchment of memory traces (Integrated Activation Patterns, or IAPs) in the lexicon by strengthening weights over temporal connections; (ii) the vertical dimension characterises parallel processing activation, and controls the number of similar, paradigmatically-related word forms that get co-activated when a member of a paradigm is input to the map (Pirrelli et al. 2014).

The two dimensions appear to pull memory organisation in opposite directions, namely serial specialisation and parallel co-activation, whose interaction accounts for interesting processing effects. When a high-frequency word is presented to a TSOM, its IAP suffers less from the competition of formally-related lower-frequency IAPs, since levels of activation of its nodes are on average higher, and inter-node connections are stronger than those in lower-frequency IAPs. This means that high-frequency words are less confusable, more fully predictable, and can be recalled more easily than their low-frequency neighbours.

On the other hand, blended patterns play an important role in the acquisition of regular and sub-regular paradigms. Verb forms sharing the same stem tend to activate partially overlapping or 'blended' IAPs. Each time any of those forms is shown to the map, the connections between shared BMUs are strengthened over again. This prompts a boosting effect in acquisition, whereby a shared stem in a paradigm is responded to by a pattern of nodes whose level of entrenchment depends on stem family frequency rather than on word token frequency. This dynamic provides an algorithmic account of the observation that regularity favours acquisition of both high- and low-frequency words, due to the facilitatory effect of having more words that consistently activate the same pattern of shared nodes. However, activation of partially overlapping IAPs, prompts a frequency-based competition for suffix selection.

Co-activation and competition are known to account for effects of family size and frequency of neighbouring words on a variety of word processing tasks. A large number of neighbours is known to support visual word recognition, with printed words with many neighbours being recognised more quickly than words with fewer neighbours (see Andrews 1997 for a review). However, when neighbours are considerably more frequent than the target word, they appear to exert an inhibitory effect on recognition of the latter. A reversed effect

from facilitation to inhibition was shown in spoken word recognition and other related tasks (Luce & Pisoni 1998; Magnuson et al. 2007), where many neighbours are found to delay recognition of a target word. Reversal from neighbour facilitation to inhibition has recently been interpreted (Chen & Mirman 2012) as an effect of parallel vs. serial input. For example, in a task of word production or written word recognition, many neighbours have a facilitatory/supporting role, and production (or written word recognition) is faster for a word in a dense neighbourhood than for an isolated word. When the input is presented serially (e.g. in spoken word recognition), high-frequency neighbours engage in a competition and exert an inhibitory effect on word processing.

We see, here, a potential connection between these effects and the frequency-by-regularity interaction in word acquisition by TSOMs. Evidence that TSOMs find it easier to acquire words when their frequency distributions are more evenly spread within their paradigms, as opposed to a situation where few members of the paradigm are more frequent than the remaining members, seems to comply with a pattern of reversal from facilitation to inhibition based on neighbour competition. We suggest accounting for this evidence in terms of a co-activation/competition dynamic in the tasks of serial word access, based on prediction, and word recall, based on parallel activation of target BMUs and filtering. When a word is input to the map, it typically co-activates other members of its own paradigm, depending on the amount of regularity in the paradigm and the frequency of paradigm members sharing the same stem. Both access and recall of the input word are thus affected by co-activation of other IAPs, but in different ways, depending on the number of forms sharing the same stem (stem family size or neighbourhood size) and their cumulative frequency (or stem family frequency). In particular, we expect a highly entropic stem family to make, on average, full word prediction (suffix prediction) of a member more difficult. Conversely, a highly entropic family makes recall easier (i.e. requiring lower filtering levels), since there is no other expected family member but the target word to be recalled from its IAP.

Accounting for the dynamic of word access/recall and the pace of word acquisition by a TSOM through the same set of memory principles that account for neighbourhood density and frequency effects in other processing tasks would show that a single computational framework can bring a number of apparently diverse effects to underlying unity. With this purpose in mind, we monitored the behaviour of Italian and German TSOMs (Section 4.2) on two tasks: serial word

access (Section 3.2.1) and word recall (Section 3.2.2). Word frequencies and morphological regularity proved to be significant predictors both in German and Italian. To use morphological regularity as a predictor, we quantified the degree of inflectional regularity of a target verb form as the number of verb forms sharing the same stem with the target (or stem family size of the target word). In fully regular paradigms, where all inflected forms share the same stem, the stem family size is equal to the paradigm size (see Appendix).

4.3.1. Task one: serial word access

In the first task, serial word access, we modelled how well an input word can be predicted by a TSOM. The more input symbols are anticipated, the easier the prediction of the target word is. As a general trend, high-frequency words are predicted more easily than low-frequency words are. In particular, a linear mixed effect model (LME)[5] of mean word prediction on German data (Figure 6), with word frequency and degree of regularity as fixed effects, shows that, for frequencies being equal, words in larger stem families are, on average, easier to be processed (and accessed) serially than words in smaller stem families. This is a type-token frequency effect: in regular paradigms (stem families[6] with 9 and 8 neighbours), and irregular ones with only a few alternant stems (stem families with 7, 6, and 5 neighbours), stem-sharing word types additively amplify the effect of token frequency on pattern consolidation, whereas words in highly irregular paradigms are typically more isolated, and thus rely more heavily (in case of suppletive forms, exclusively) on their own token frequency.

In addition, more regular word forms can benefit from longer stems (Figure A.3 in Appendix), thus increasing the average number of predicted input symbols.

On the other hand, co-activation of stem sharing words triggers competition for suffix selection. In words with larger stem families, inflectional endings are more difficult to predict than in words with smaller stem families. On average, suffixes are better predicted (i.e. they are more easily predicted by their stems) in more irregular verb families than in regular paradigms, where a greater number of neighbours sharing the same stem selects different inflectional endings.

The more verb forms share the same stem, the stronger the competition for accessing an inflectional suffix, as confirmed by a second linear model fitting German suffix prediction only, with word frequency, degrees of stem regularity, and suffix length as fixed effects (Figure 7, left). Suffixes in words that are surrounded by more competitors are less easy to predict when they are in the

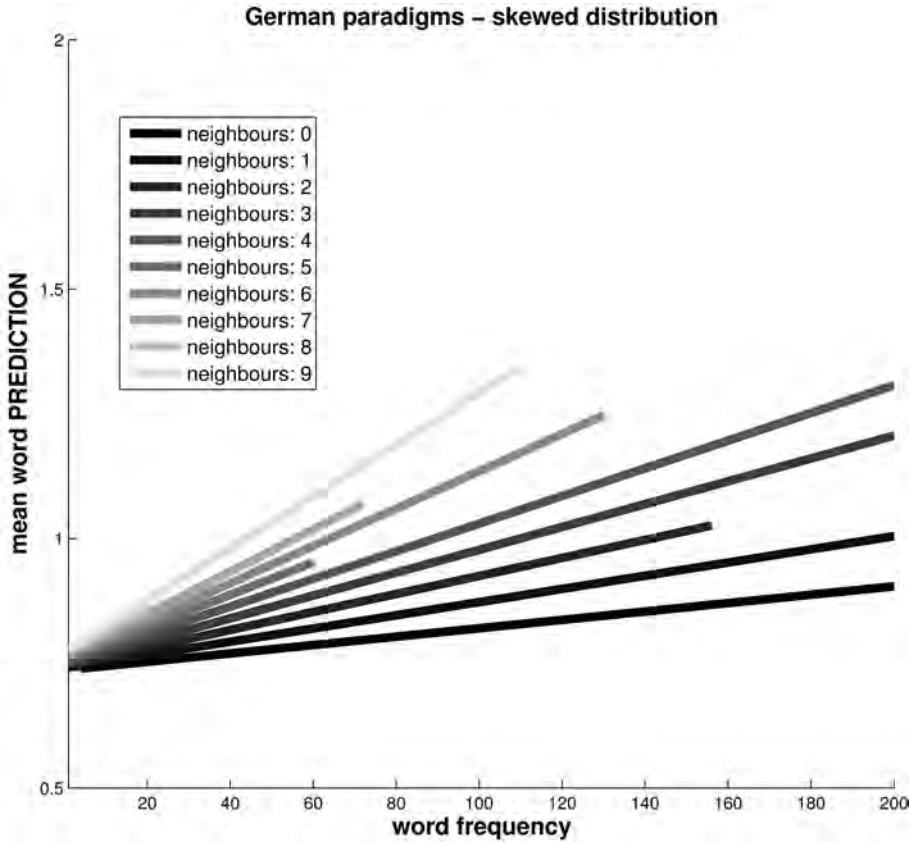

Figure 6. Marginal plot of interaction effects between word frequency (*x*-axis) and degrees of stem regularity (number of neighbours) in an LME model fitting mean word prediction (*y*-axis) by TSOMs trained on German verb forms in the SD regime. Random effects: TSOM instances (n = 5), paradigms (n = 50). Fixed effects: word frequency, number of neighbours.

low-medium frequency range. We observe, however, that the facilitatory effect of increasing frequency has a steeper rate when words are embedded in bigger stem families, where the few regulars in the high frequency range benefit both from their own token frequency and from absence of highly frequent competitors. In fact, a small increase in token frequency provides regulars with a comparatively larger advantage in suffix prediction, since regular paradigms present on average higher levels of stem family entropy than irregulars do (Figure A.3 in Appendix).

The same LME model fitted to Italian data (Figure 7, right) confirms that an increase in word frequency mostly favours prediction of suffixes in bigger stem families. The interaction of word frequency and degrees of stem regularity is even clearer in Italian, where suffixes are on average longer than in German (Figure A.2 in Appendix).

The three models confirm that word forms containing recurrent sub-lexical structures can take advantage of the memory traces shared by other related forms: connections between shared nodes in 'blended' IAPs are strengthened more quickly since recurrent sub-lexical structures are shown more often in training, similarly to what happens with high-frequency isolated words. Conversely, stem sharing increases the amount of uncertainty in the selection of an upcoming inflectional ending, as a function of the cardinality of the stem family and the frequency distribution of its members.

This dynamic interaction can be observed in more detail by comparing average symbol prediction across different stem and suffix positions for the two languages: German and Italian verb forms (Figure 8 top and bottom, respectively) in the uniform (left) and skewed (right) distributions.

In the training regime with no token frequency effects (Figure 8, left plots), where inflected forms are uniformly distributed, acquisition of regulars is typically paradigm-based, and regulars are, on average, more easily predicted than irregulars, as an effect of type

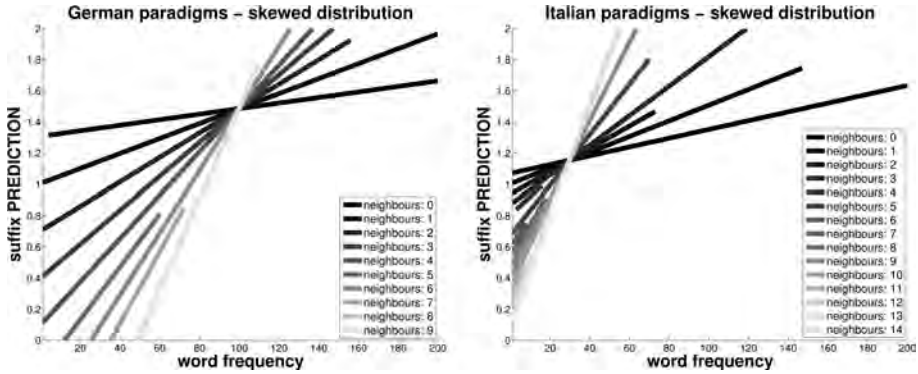

Figure 7. Marginal plot of interaction effects between word frequency (x-axis) and degrees of stem regularity (number of neighbours) in an LME model fitting suffix net prediction (y-axis) by TSOMs trained on German (left) and Italian (right) verb forms in the SD regime. Random effects: TSOM instances (n = 5), paradigms (n = 50). Fixed effects: word frequency, number of neighbours, suffix length.

frequency and overlaying of redundant morphological patterns (as shown in Figure 6). This is confirmed by the higher prediction rate for regular stems compared with irregular stems in both languages, shown by the steeper dashed lines across symbol positions in the stem (-6:-1 range on the *x*-axis, Figure 8, left plots), where the '-1' position identifies the end of stems and '0' the suffix onset.

It is important to emphasise that, in both languages, regularity increases the amount of processing uncertainty in predicting a suffix, as shown by the drop in prediction at morpheme boundary, and by the different slopes of dashed and solid lines across symbol positions in

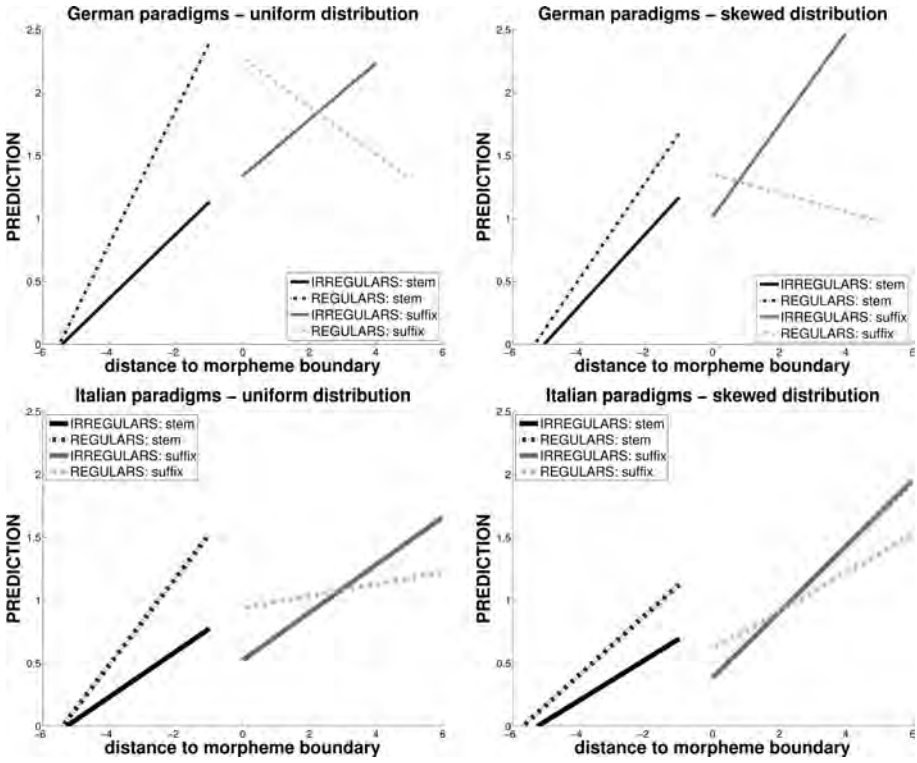

Figure 8. Marginal plots of interaction effects between symbol distance to the morpheme-boundary (*x*-axis) and stems/suffixes in regular (dashed lines) and irregular (solid lines) paradigms, in an LME model fitting symbol prediction (*y*-axis) by TSOMs trained on German (top) and Italian (bottom) verb forms in both UD (left) and SD (right) regimes. Random effects: TSOM instances (n = 5), word forms. Fixed effects: distance to morpheme boundary, regular/irregular paradigm, stem/suffix, word frequency, suffix length.

the suffix (0:6 range on the *x*-axis). This evidence is differently modulated by real frequency distributions (Figure 8, right plots), where stems in regular paradigms are predicted less easily (less steeply ascending lines), due to a greater amount of memory resources devoted to highly frequent words (most of which are irregulars).[7]

It is less easy to fully predict endings in German regulars, as a consequence of the highly embedded structure of German inflectional markers, illustrated by the word graph representation of Figure 3 (Section 3.2). Branching nodes, in fact, cause uncertainty in serial processing and increase competition for suffix selection. In irregular paradigms, stem alternation considerably reduces the number of endings.

Italian shows a similar pattern, with inflectional markers being longer and more distinguishable at earlier positions in both irregulars and regulars.

4.3.2. Task two: word recall

Turning to the second task, ease of word recall from IAPs is measured in terms of the amount of filtering required to accurately recall a word from its IAP (Section 3.2.2), with easy-to-recall words taking little or no filtering to be recalled accurately. We fitted a linear mixed effect model of word filtering with word token frequency and stem family size of the target word as fixed effects, for both uniform and skewed training regimes, whose marginal plots are shown in Figure 9.

When verb forms are uniformly distributed (Figure 9, left), regular and sub-regular paradigms (i.e. paradigms with a larger stem family size) are easier to recall: they require lower filtering levels, because members of the same stem family tend to develop blended activation patterns and benefit from cumulative activation of more word types sharing the same stem. In fact unlike serial word access, which only relies on the map's prediction bias for the most likely candidate symbol in a pool of competing candidates (Equation 3), recall is based on the integration of the map's temporal expectation (y_T) with the IAP ($\hat{y}$) of the target word to recall (Equation 4).

When we move to more realistic distributions (Figure 9 right plot), we observe an interesting frequency-by-regularity interaction. In the medium-high frequency range, words in more irregular paradigms are more easily recalled than regulars with the same word frequency. This is a consequence of the comparatively differential advantage that type and token frequencies give to irregulars and regulars. Irregulars benefit more from an increase in word frequency

than regulars do, since an increase in token frequency makes internode connections stronger, levels of activation higher, non-target word co-activation lower. This dynamic directly affects ease of recall, in particular for words in irregular paradigms, whose IAPs contain fewer if any co-activated neighbours.

With regulars, neighbour co-activation is stronger and it weighs down the potential contribution of token frequency to facilitating recall. Conversely, in the low-frequency range, competition in regular paradigms is thus easily offset by the larger facilitatory contribution of many overlaying IAPs.

To focus on effects of co-activation/competition between blended IAPs within densely populated neighbour families (regular paradigms), we fitted an LME model to suffix filtering in regular paradigms only, using word frequency, stem family entropy and stem family frequency as fixed effects (Figure 10). More entropic stem families (i.e. families with uniformly distributed members) make it comparatively easier for words in the low-frequency range to be recalled than less entropic stem families do. Words in less entropic stem families require more suffix filtering in the low-frequency range, since they suffer the competition of high-frequency neighbours, but get more facilitation as word frequency grows. The interaction illustrates the impact of frequency distributions on competition for suffix recall. When a target word belongs to a highly entropic stem family, the facilitatory impact on suffix recall increases

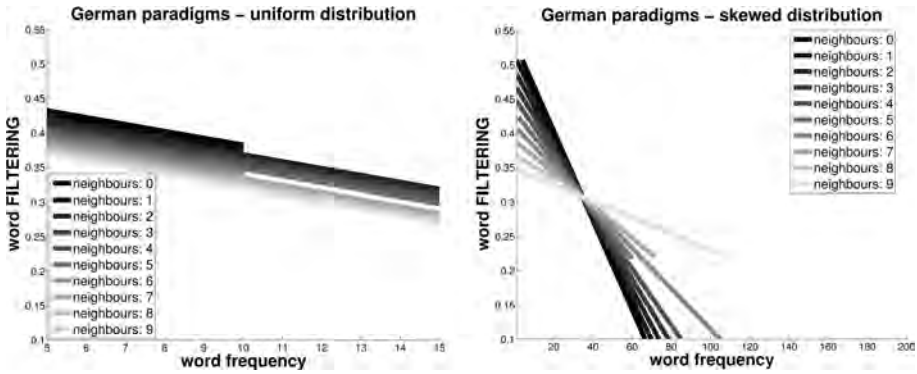

Figure 9. Marginal plot of interaction effects between word frequency (*x*-axis) and degrees of stem regularity (number of neighbours) in an LME model fitting mean word filtering (*y*-axis) by TSOMs trained on German verb forms in both UD (left) and SD (right) regimes. Random effects: TSOM instances (n = 5), paradigms (n = 50). Fixed effects: word frequency, number of neighbours, stem family entropy, stem family frequency.

only marginally with frequency. Conversely, in a low-entropy family, a low-frequency member suffers from the competition of higher frequency members: by increasing word frequency, we are simply shifting our focus on the strongest competitors, whose recall gets increasingly easier as family entropy goes down.

To sum up, effects of frequency and regularity are the result of the interaction of a common pool of principles of correlative learning, but they are dependent on both training regime and processing task. Regularity is based on larger stem families (and relies on higher type frequencies), thus compensating for lower token frequencies with the joint support of family members. High token frequency, on the other hand, favours entrenchment of individual items, although it tends to interfere in families with low frequency members.

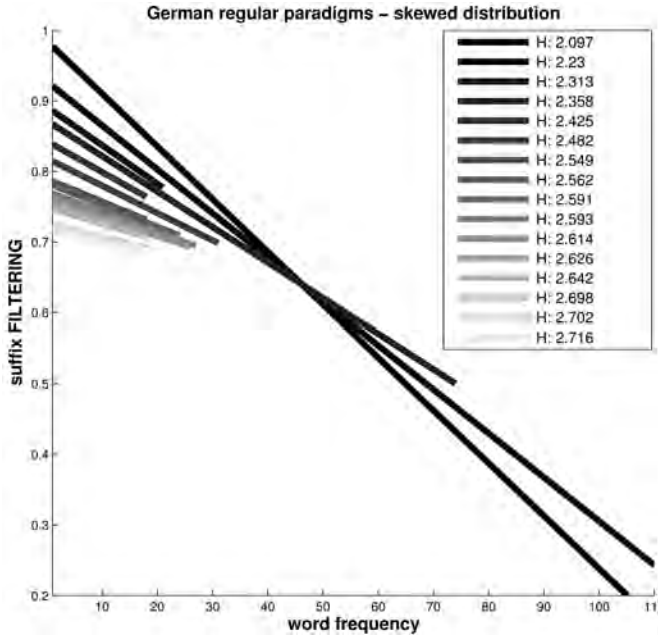

Figure 10. Marginal plot of interaction effects between word frequency (*x*-axis) and levels of stem family entropy (H) in an LME model fitting mean suffix filtering (*y*-axis) by TSOMs trained on German regular verb forms in the SD regime. Random effects: TSOM instances (n = 5), paradigms (n = 16). Fixed effects: word frequency, stem family entropy, stem family frequency.

5. General Discussion

TSOMs offer a highly redundant model of lexical memory, where specialised IAPs, responding to a few input forms only, coexist with many 'blended' (or less specialised) IAPs, meeting the input requirements of several members of the same family, whose acquisition can take advantage of cumulative stem frequencies.

The degree of specialisation of a pattern is defined by the number of strong temporal connections departing from each BMU in the pattern. The fewer these connections are, the more specialised the pattern is. It has been observed (Ferro et al. 2010b) that pattern specialisation is advantageous for word processing, since dedicated memory connections minimize the number of one-to-many inter-node transitions (Figure 2, right), thus reducing the degree of uncertainty in accessing and recalling a word form stored in a TSOM. The quantitative analysis offered in this paper supports this claim through a broader range of empirical evidence. In TSOMs, long-term expectations enhance successful prediction of upcoming symbols, and make it easier for a map to recall a sequence through its IAP.

This highlights one of the most distinctive features of integrative models of lexical memory. Since long-term expectations are based on the probabilistic distribution of successful processing strategies, stored lexical representations can be conceived of as routinized processing patterns. These patterns are ready-made processing routines that the lexical processor can flexibly use on demand, depending on input requirements. The evidence reported in this paper helped us understand more about the factors affecting this dynamic integration of representations and processing responses. The influence of these factors may vary as a function of the processing task.

5.1. Acquisition

The amount of general vs. specialised resources that are apportioned by a TSOM through learning largely correlates with sensitivity to a gradient of morphological regularity, and makes contact with differences in processing strategies (serial processing vs. parallel activation).

More regular forms share redundant structures with other words to a considerable extent: they are stored in blended activation patterns and processed accordingly. Their acquisition is ultimately a function of how often these shared structures are found in input. Thus the effect of word token frequency on entrenchment can capitalize on the cumulative token frequency of all members of the same

word family, whose contribution is additive. In the end, having more neighbours to rely on favours acquisition (Figure 5).

Highly irregular forms are, by definition, relatively isolated, and get little (if any) support from the overall organisation of the majority of lexical forms. Hence, they tend to develop IAPs that are rarely used elsewhere by a TSOM. Although this may offer a competitive advantage in early acquisition, where high-frequency words are acquired first (Figure 4), it is arguably not the most effective strategy for the acquisition of lexical data, which are attested with a Zipfian distribution. Since the level of memory entrenchment of highly irregulars chiefly depends on how often they appear in the input (their token frequency), acquisition can only rely on rote learning, rather than on associative relations with already stored items.

5.2. Serial processing and lexical access

Serial processing of an input word crucially exploits the predictive power of forward temporal connections in IAPs. When we control for word frequency, forms in more regular paradigms are, on average, quicker to be processed serially than forms in irregular paradigms (Figure 6), since the former tend to cluster in larger word families, and this makes regulars more familiar or 'wordlike', and their blended IAPs more 'routinized'. However, stem sharing increases the amount of uncertainty in the selection of an upcoming suffix (Figure 7), accounting for effects of slower completion of words with more neighbours. This is a function of the cardinality of the word family, and of frequency distribution of family members. In our training sets, word token frequency was distributed more uniformly in regular paradigms than in irregular paradigms (see Appendix), and this increased the amount of processing uncertainty at the stem-suffix boundary in regular paradigms.

On the contrary, forms in more irregular paradigms, suffer less from interference caused by co-activation of overlapping IAPs, and due to their higher frequency they develop more deeply entrenched temporal connections. Furthermore, irregularly inflected words typically undergo stem alternation processes (as in *finden* vs. *fanden*, or *vengo* vs. *vieni*), which bring forward their recognition uniqueness point, and reduce competition for suffix selection.

Although it is always difficult to draw general conclusions from corpus-based frequency distributions, we believe this to be a general pattern in the inflectional morphology of the two languages. It is a remarkable fact that TSOMs prove to be sensitive to structure-based effects at morpheme boundaries, and that these effects correlate with

measurable levels of weight distribution over connections straddling morpheme boundaries (Figure 8). This evidence shows that paradigm entropy is in fact a measure of how evenly members of the same word family compete/co-activate in processing. Uniform distributions prompt tighter competition for suffix selection and this is, in general, detrimental for serial word access.

5.3. Parallel activation and word recall

Does parallel activation help word recall? We observed that it does, but with some qualifications. The facilitative effect on word recall of highly entropic word families is particularly prominent when the frequency distribution of family members is in the low-medium range (Figure 9, right). For forms with comparatively low frequency, there is an advantage in being surrounded by many family members. Shared stem patterns largely offsets potential competition for suffix recall. In recalling a word from its IAP, a TSOM uses both its own temporal expectations (which may be conflicting if more words are co-activated) and contextual information available in the target IAP (which helps resolve conflicts). However, this is the result of a dynamic balance, which can be tipped off by an increase in token frequency of family members. In the high-frequency range, irregulars are in fact recalled increasingly more easily, while regulars becoming comparatively less advantaged. This dynamic explains two important trends in our sets of experimental evidence on the pace of word/paradigm acquisition. First, high-frequency words are learned more quickly, but their low-frequency family companions are learned less easily. What is an advantage for item-based acquisition turns out to be a hurdle for paradigm-based acquisition. More regular items, on the other hand, tend to develop blended activation patterns, which benefit from the cumulative frequency of word family members. Blended IAPs are good for generalisation in acquisition, as they can be used for transferring knowledge of some members of the family to the whole family. This explains why more regular paradigms are acquired more quickly than irregular paradigms and are less affected by effects of token frequency distributions. At the same time, however, they suffer under competition in a task of serial word completion to a greater extent.

5.4. Concluding remarks

Our simulative evidence accords well with competition-based and usage-based models of language acquisition (Tomasello 2003; MacWhinney 2008), making the further suggestion that integrative models of memory self-organisation can account for the appar-

ent dualism between item-based acquisition of irregular forms and paradigm-based acquisition of regulars. We observed that, in TSOMs, item-based acquisition of high-frequency irregulars plays a major role at early learning epochs, as some irregular items are very frequent and develop dedicated IAPs. Type frequency effects emerge only later, due to the overlaying of redundant morphological patterns, but they play an increasingly important role in an emergent lexicon, shifting acquisitional strategies from word rote memorisation to dynamic memory-based generalisation. This general trend is influenced by degrees of morphological regularity and by the interaction between frequency and regularity, with frequency speeding up word acquisition, and regularity speeding up paradigm acquisition.

It should be emphasised that the present account of the frequency-by-regularity interaction in word and paradigm acquisition is in line with general principles of memory self-organisation, and with effects of neighbour family size and frequency on word processing. We believe this convergence to be neither accidental nor trivial. One of the goals of our simulations was to show that a single computational framework, with the same set of parameters, can account for (i) effects of processing and memory interaction, (ii) differential effects of frequency and gradients of regularity in different processing tasks (e.g. serial lexical access vs. word recall), and (iii) facilitation to inhibition reversal within the same task, depending on the interaction between regularity and frequency. The final goal was to bring several contrasting effects, which have so far been analysed and accounted for in terms of different computational models (with the exception of Chen & Mirman 2012), to underlying unity. In addition, proof that effects of morphological redundancy are the specific, emergent outcome of more general, pre-morphological principles of memorisation of symbolic time-series, may also have interesting theoretical implications on issues of lexical architecture, suggesting that pre- and post-lexical effects of word priming can in fact be based on a common pool of underlying processing mechanisms.

Appendix

Data in the two training sets show interesting common patterns as well as differences between German and Italian. Figure A.1 shows the box plot distribution of the stem family size for each paradigm in the training sets for German and Italian. A stem family is defined as the set of paradigmatically-related forms inflected on the basis of the same stem (as in German *machen*, *mache*, *macht*, *machen*, *machten* etc.). In regular paradigms, all inflected forms select a unique stem form (typically the infinitive stem or the present indicative stem), which may undergo systematic processes of stem formation in predictable paradigm cells (e.g. *mach-en* 'make' vs. *ge-mach-t* 'made' past participle). In irregular paradigms inflected forms may show unsystematic patterns of stem variation across possibly unpredictable cells (e.g. German *denk-en* 'think' vs. *ge-dach-t* 'thought' past participle, and Italian *venire* 'come' vs. *veng-o*, *vien-i* 'come' first and second person singular, present indicative). This makes it more difficult for a speaker to predict unknown inflected forms in highly irregular paradigms than in regular and sub-regular paradigms. Hence, an estimate of how many different stem families are attested in a single paradigm, and how many members each family has, defines a gradient of paradigmatic regularity, and, ultimately, gives information on how difficult a paradigm is to learn.

Both training samples contain more instances of irregular paradigms (i.e. paradigms requiring more than one stem for its inflected forms) than regular ones, the former occurring on average significantly more often than the latter. Variance of stem family size is wider in Italian paradigms than in German ones. This confirms that the overall organisation of irregular Italian paradigms tends to be more fragmented and less predictable, with irregular paradigms containing more scattered stem families of significantly different size.

The two data sets are largely comparable on mean word frequency and mean stem length (see radar plot in Figure A.2), but inflectional endings are, on average, longer in Italian than in German. Furthermore, because we count family members by the number of non-homographic forms attested in each paradigm, Italian stem families appear to be larger (i.e. with a greater number of stem family neighbours, or NNB) than German stem families, where, for example, the infinitive form (e.g. *machen*) is also found in two more paradigm cells. Likewise, members of Italian stem families are more evenly distributed than members of German stem families, as shown by their larger entropy scores. Finally, both languages

exhibit comparable differences between regulars and irregulars in terms of stem length, stem family size, and entropy of the stem family (Figures A.3 left and right).

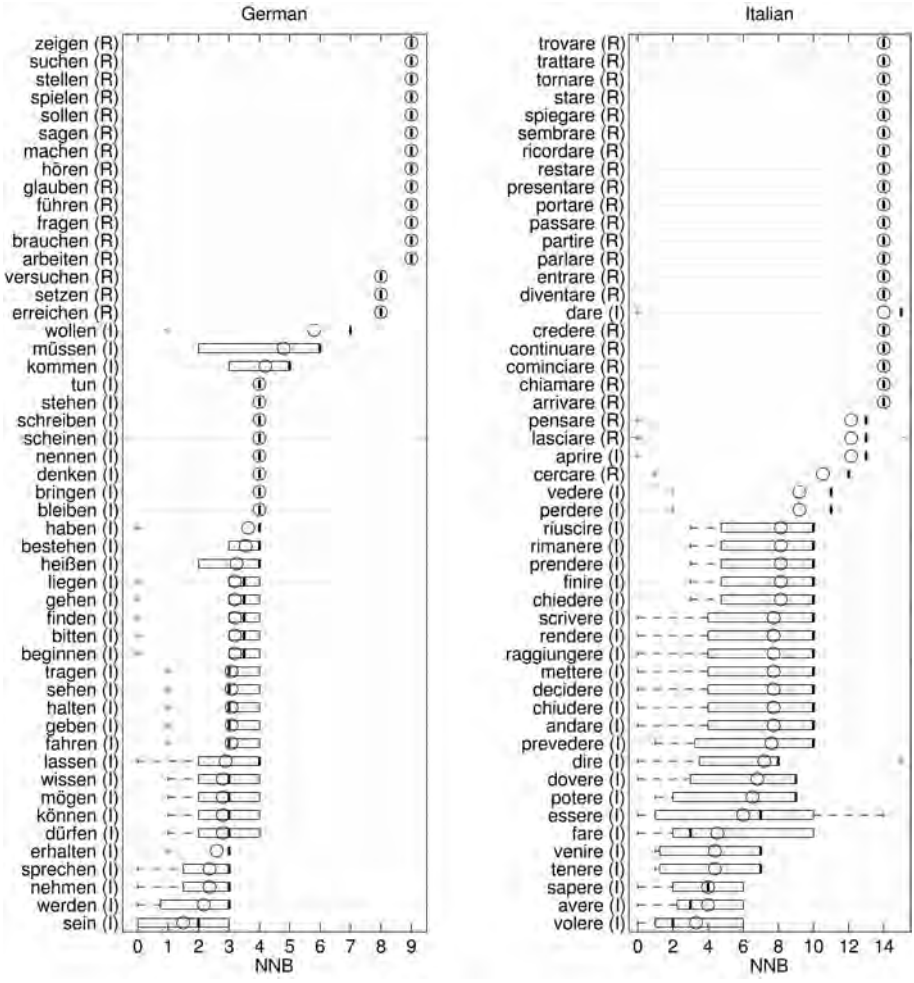

Figure A.1. Box plot distribution of stem family size for German (Left) and Italian (Right) paradigms in the two training sets. Paradigms are ordered top-down by decreasing values of paradigmatic regularity, namely by the mean number of stem family members (NNB, number of neighbours). Circles mark the family size mean, and bold lines the family size median. Labels in brackets stand for 'regular' (R) and 'irregular' (I), according to a traditional dichotomous classification. '+' signs mark data outliers.

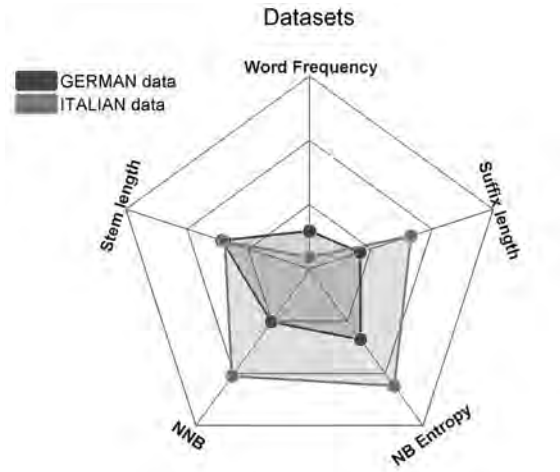

Figure A.2. A radar plot of mean values for word frequency, stem and suffix length, stem family size (NNB: number of stem family neighbours) and stem family entropy (NB Entropy), in German and Italian training sets.

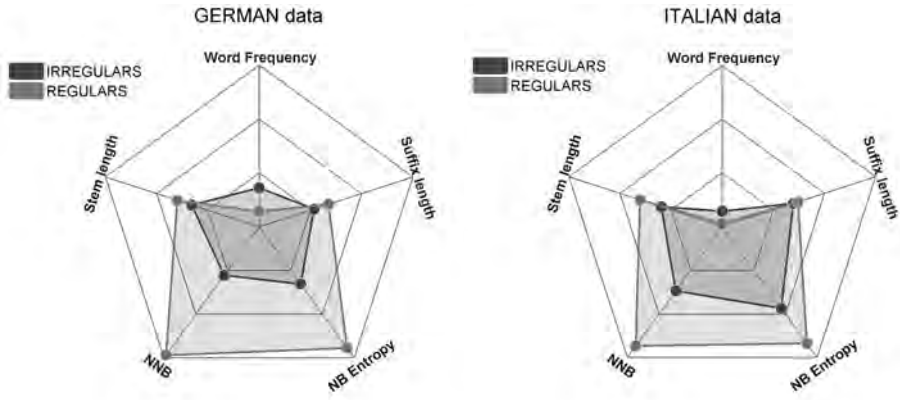

Figure A.3. Radar plots of mean values for word frequency, stem and suffix length, stem family size (NNB: number of stem family neighbours) and stem family entropy (NB Entropy) in regular and irregular paradigms, for German (Right) and Italian (Left) training sets.

Notes

[1] Recall scores at the end of training are as follows: UD Italian (99.7%, std=0.1%), SD Italian (types: 99.4%, std=0.3%; tokens: 99.9%, std=0.1%), UD German (99.8%, std=0.1%), SD German (types: 99.7%, std=0.2%; tokens: 99.9%, std=0.1%).

[2] In the whole paper, the two-sided Wilcoxon rank sum test is used to test the null hypothesis that data contain samples from continuous distributions with equal medians, against the alternative that they are not. The test assumes that the two samples, which can be of different length and whose distribution can be not normal, are independent.

[3] Following Aronoff (1994) and Pirrelli (2000), inflectional (ir)regularity is defined here as a function of the number of unpredictable stem formation processes that apply within a verb paradigm. In fully regular paradigms, inflected forms require a single stem, which may undergo systematic changes in predictable paradigm cells (e.g. *mach-en* 'make' vs. *ge-mach*-t 'made' past participle). In irregular paradigms, inflected forms present unsystematic patterns of stem variation across possibly unpredictable cells (e.g. *denk-en* 'think' vs. *ge-dach-t* 'thought' past participle). This traditional dichotomous classification can be made more gradient by classifying verb paradigms according to the number of unpredictable stems they require (see Appendix).

[4] The Pearson product-moment correlation coefficient is calculated as a measure of the degree of linear dependence between two variables, giving a value between +1 and -1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation. The two variables are supposed to be normally distributed.

[5] All linear plots in the paper are marginal plots of Linear Mixed Effects (LME) models, relating some characteristics of the training dataset (predictors) to the response of a TSOM (dependent variable). Predictors differ across models depending on the dependent variable and the theoretical questions being addressed. Plots are obtained ignoring the contribution to the model response coming from predictors that are considered as random effects: namely, the specific trained instance of the TSOM, the items in the dataset, and, when explicitly stated, also their paradigms. Only coefficients that are associated with fixed effects are used. When a fixed effect predictor is not directly shown in a model graph, the plot is obtained by using the predictor's average value over the entire dataset. Unless stated otherwise, the range of each line in the graphs (i.e. the range of the predictor used on the x-axis) corresponds to the range of the actual data in the dataset.

[6] A stem family is defined as a set of paradigmatically-related forms whose inflection requires the same stem (see note 3 and Appendix). In regular paradigms, all inflected forms belong to the same stem family. Irregular paradigms have more stem families, depending on their level of irregularity. For each inflected form, we can count the number of inflected forms (neighbours) that are members of the same stem family: the more the neighbours, the more regular and predictable the inflected form.

[7] The Pearson correlation, calculated between the token frequency of a word and the average number of BMUs responding only to that word, is r=.42, p-value <.00001 for German, and r=.34, p-value <.00001 for Italian.

Bibliographical References

Ackerman, Farrell; Blevins, James & Malouf, Robert 2009. Parts and wholes: implicative patterns in inflectional paradigms. In Blevins, James P. & Blevins, Juliette (eds.), *Analogy in Grammar*. Oxford: Oxford University Press. 54-82.

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

Alegre, Maria & Gordon, Peter 1999. Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40. 41-61.

Andrews, Sally 1997. The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Revie*w 4. 439–461.

Aronoff, Mark 1994. *Morphology by itself: stems and inflectional classes*. Cambridge: MIT Press.

Baayen, Harald R. 2007. Storage and computation in the mental lexicon. In Jarema, Gonia & Libben, Gary (eds.), *The Mental Lexicon: Core Perspectives*. Amsterdam: Elsevier. 81-104.

Baayen, Harald R. & Schreuder, Robert 2000. The morphological family size effect and morphology. *Language and Cognitive Processes* 15. 329-365.

Baayen, Harald R. & Schreuder, Robert 1999. War and peace: Morphemes and full forms in a non-interactive activation parallel dual-route model. *Brain and Language* 68. 27-32.

Baayen, Harald R.; Lieber, Rochelle & Schreuder, Robert 1997. The morphological complexity of simplex nouns. *Linguistics* 35. 861-877.

Baayen, Harald R.; Piepenbrock, Richard & Gulikers, Leon 1995. The CELEX Lexical Database (CD-ROM). Philadelphia: Linguistic Data Consortium.

Bailey, Todd M. & Hahn, Ulrike 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44. 568-591.

Bittner, Dagmar; Dressler, Wolfgang U. & Kilani-Schoch, Marianne (eds.), 2003. *Development of Verb Inflection in First Language Acquisition: a cross-linguistic perspective*. Berlin: De Gruyter Mouton.

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

Burzio, Luigi 2004. Paradigmatic and syntagmatic relations in Italian verbal inflection. In Auger, Julie; Clements, Clancy J. & Vance, Barbara (eds.), *Contemporary Approaches to Romance Linguistics*. Amsterdam: John Benjamins. 17-44.

Bybee, Joan 1995. Regular Morphology and the Lexicon. *Language and Cognitive Processes* 10. 425-455.

Bybee, Joan & Moder, Carol L. 1983. Morphological Classes as Natural Categories. *Language* 9. 251-270.

Bybee, Joan & Slobin, Dan I. 1982. Rules and Schemas in the development and use of the English Past Tense. *Language* 58. 265-289.

Catani, Marco; Jones, Derek K. & Ffytche, Dominic H. 2005. Perisylvian language networks of the human brain. *Annals of Neurology* 57. 8-16.

Chen, Qi & Mirman, Daniel 2012. Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review* 119. 417-430.

Colombo, Lucia; Laudanna, Alessandro; De Martino, Maria & Brivio, Cristina 2004. Regularity and/or consistency in the production of the past participle? *Brain and Language* 90. 128-142.

D'Esposito, Mark 2007. From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society B. Biological Sciences* 362. 761-772.

Dabrowska, Ewa 2005. Productivity and beyond: mastering the Polish genitive inflection. *Journal of Child Language* 32. 191-205.

Dabrowska, Ewa 2004. Rules or schemata? Evidence from Polish. *Language and Cognitive Processes* 19. 225-271.

Daelemans, Walter & van den Bosch, Antal 2005. *Memory-based language processing*. Cambridge UK: Cambridge University Press.

Ferro, Marcello; Pezzulo, Giovanni & Pirrelli, Vito 2010a. Morphology, Memory and the Mental Lexicon. In Pirrelli, V. (ed.), *Lingue e Linguaggio* IX. 199-238.

Ferro, Marcello; Ognibene, Dimitri; Pezzulo, Giovanni & Pirrelli, Vito 2010b. Reading as active sensing: A computational model of gaze planning in word recognition. *Frontiers in Neurorobotics* 4. 1-16.

Finkel, Raphael & Stump, Gregory 2007. Principal parts and morphological typology. *Morphology* 17. 39-75.

Ford, Michael A.; Marslen-Wilson, William D. & Davis, Matthew H. 2003. Morphology and frequency: contrasting methodologies. In Baayen, Harald R. & Schreuder, Robert (eds.), *Morpho-logical Structure in Language Processing*. Berlin / New York: De Gruyter Mouton. 89-124.

Forster, Kenneth. I. 1976. Accessing the mental lexicon. In Wales, Roger J. & Walker, Edward (eds.), *New Approaches to Language Mechanisms*. Amsterdam: North-Holland. 257-287.

Gaskell, M. Gareth & Marslen-Wilson, William D. 2002. Representation and competition in the perception of spoken words. *Cognitive Psychology* 45. 220–266.

Gathercole, Susan E.; Hitch, Grahama J.; Service, Elisabet S. & Martin, Amanda J. 1997. Phonological short term memory and new word learning in children. *Developmental Psychology* 33. 966-979.

Kohonen, Teuvo 2001. *Self-organizing maps*. Berlin Heidelberg: Springer-Verlag.

Koutnik, Jan 2007. Inductive Modelling of Temporal Sequences by Means of Self-organization. In *Proceeding of International Workshop on Inductive Modelling*. Prague. 269-277.

Levelt, Willelm J. M., Roelofs, Ardi & Meyer, Antje S. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22. 1–75.

Luce, Paul A. 1986. A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics* 39. 155–158.

Luce, Paul A. & Pisoni, David B. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19. 1-36.

Lüdeling, Anke & de Jong, Nivja 2002. German particle verbs and word for-

mation. In Dehé, Nicole; Jackendoff, Ray; McIntyre, Andrew & Urban Silke (eds.), *Explorations in Verb-Particle Constructions*. Berlin: De Gruyter Mouton.

Lyding, Verena; Stemle, Egon; Borghetti, Claudia; Brunello, Marco; Castagnoli, Sara; Dell'Orletta, Felice; Dittmann, Henrik; Lenci, Alessandro & Pirrelli, Vito 2014. The PAISÀ Corpus of Italian Web Texts. In Bildhauer, Felix & Schäfer, Roland (eds.), *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*. Gothenburg. 36-43.

Ma, Wei J.; Husain, Masud & Bays, Paul M. 2014. Changing concepts of working memory. *Nature Neuroscience* 17. 347-356.

MacWhinney, Brian 2008. A unified model of language acquisition. In Kroll, Judith F. & De Groot, Annette M.B. (eds.), *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford: Oxford University Press. 49-67.

MacWhinney, Brian & Leinbach, Jared 1991. Implementations are not conceptualizations: Revising the verb learning model. *Cognition* 40. 121-157.

Magnuson, James S.; Dixon, James A.; Tanenhaus, Michael K. & Aslin, Richard N. 2007. The dynamics of lexical competition during spoken word recognition. *Cognitive Science* 31. 1-24.

Marslen-Wilson, William D. 1993. Issues of process and representation in lexical access. In Altmann, Gerry & Shillcock, Richard (eds.), *Cognitive Models of Speech Processes: The Second Sperlonga Meeting*. Hove: Lawrence Erlbaum Associates. 187-210.

Marzi, Claudia 2014. Models and dynamics of the morphological lexicon in mono- and bilingual acquisition. *Unpublished PhD Dissertation*. University of Pavia.

Marzi, Claudia & Pirrelli, Vito 2015. A Neuro-Computational Approach to Understanding the Mental Lexicon. *Journal of Cognitive Science* 16. 491-533.

Marzi, Claudia; Ferro, Marcello & Pirrelli, Vito 2014. Morphological structure through lexical parsability. *Lingue e Linguaggio* XIII. 263-290.

Marzi, Claudia; Ferro, Marcello & Pirrelli, Vito 2012. Word alignment and paradigm induction. *Lingue e Linguaggio* XI. 251-274.

Matthews, Peter H. 1991. *Morphology (second edition)*. Cambridge, UK: Cambridge University Press.

McClelland, James L. & Elman, Jeffrey L. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18. 1-86.

Moscoso del Prado Martin, Fermin 2007. Co-occurence and the effect of inflectional paradigms. *Lingue e Linguaggio* VI. 247-262.

Moscoso del Prado Martin, Fermin; Kostić, Aleksandar & Baayen, Harald R. 2004. Putting the bits together: an information theoretical perspective on morphological processing. *Cognition* 94. 1-18.

Norris, Dennis; McQueen, James M. & Cutler, Ann 1995. Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21. 1209–28.

Orsolini, Margherita; Fanari, Rachele & Bowles, Hugo 1998. Acquiring regular and irregular inflections in a language with verb classes. *Language and Cognitive Processes* 13. 425-464.

Orsolini, Margherita & Marslen-Wilson, William 1997. Universals in morphological representation: Evidence from Italian. *Language and Cognitive Processes* 12. 1-47.

Pinker, Steven & Ullman, Michael 2002. The past and future of the past tense. *Trends in Cognitive Science* 6. 456-463.

Pirrelli, Vito 2000. *Paradigmi in morfologia. Un approccio interdisciplinare alla flessione verbale dell'italiano*. Pisa / Roma: Istituti editoriali e poligrafici internazionali.

Pirrelli, Vito; Ferro, Marcello & Marzi, Claudia 2015. Computational complexity of abstractive morphology. In Baerman, Matthew; Brown, Dustan & Corbett, Greville (eds.), *Understanding and Measuring Morphological Complexity*. Oxford: Oxford University Press. 141-166.

Pirrelli, Vito; Marzi, Claudia & Ferro, Marcello 2014. Two-dimensional Wordlikeness Effects in Lexical Organisation. In Basili, Roberto; Lenci, Alessandro & Magnini, Bernardo (eds.), *Proceedings of the First Italian Conference on Computational Linguistic*. Pisa: Pisa University Press. 301-305.

Pirrelli, Vito; Ferro, Marcello & Calderone, Basilio 2011. Learning paradigms in time and space. Computational evidence from Romance languages. In Maiden, Martin; Smith, John C.; Goldbach, Maria & Hinzelin, Marc O. (eds.), *Morphological Autonomy: Perspectives from Romance Inflectional Morphology*. Oxford: Oxford University Press. 135-157.

Pirrelli, Vito & Battista, Marco 2000. The paradigmatic dimension of stem allomorphy in Italian verb inflection. *Italian Journal of Linguistics* 12. 307-379.

Pitt, Mark A. & McQueen, James M. 1998. Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language* 39. 347-370.

Plunkett, Kim & Juola, Patrick 1999. A connectionist model of English past tense and plural morphology. *Cognitive Science* 23. 463-490.

Rumelhart, David E. & McClelland, James L. 1986. On learning the past tense of English verbs. In McClelland, James L. & Rumelhart, David E. (eds.), *Parallel distributed processing*. Cambridge: MIT Press. 217-270.

Sandra, Dominiek 1994. The morphology of the mental lexicon: Internal word structure viewed from a psycholinguistic perspective. *Language and Cognitive Processes* 9. 327-269.

Schreuder, Robert & Baayen, Harald R. 1997. How complex simplex words can be. *Journal of Memory and Language* 37. 118-139.

Shalom, Dorit B. & Poeppel, David 2008. Functional Anatomic Models of Language: Assembling the Pieces. *The Neuroscientist* 14. 119-127.

Tomasello, Michael 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Vitevitch, Michael S. & Luce, Paul A. 1998. When words compete: Levels of processing in spoken word recognition. *Psychological Science* 9. 325-329.

Vitevitch, Michael S.; Luce, Paul A.; Charles-Luce, Jan & Kemmerer, David 1997. Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech* 40. 47-62.

Wilson, Margaret 2001. The case of sensorimotor coding in working memory. *Psychonomic Bulletin and Review* 8. 44-57.