# Lexical parsability and morphological structure

Marcello Ferro, Claudia Marzi, Vito Pirrelli
*Institute for Computational Linguistics – National Research Council - G. Moruzzi 1 – Pisa, Italy*
marcello.ferro@ilc.cnr.it, claudia.marzi@ilc.cnr.it, vito.pirrelli@ilc.cnr.it

## 1. Introduction

A classical tenet in the psycholinguistic literature on the mental lexicon is that parsed affixes are associated with independently activated access units that tend to spread activation to other affix-sharing words, and that activation levels strongly correlate with the affix productivity. A number of influential papers (Hay 2001, Hay and Baayen 2002, Hay and Plag 2004) suggested that parsability criteria interact with frequency to define morphological productivity and word structure constraints in the lexicon. For example, the frequency of a derivative (e.g. *government*) relative to its base (*govern*) is shown to be a good predictor for parsability/productivity. The higher the base/derivative frequency ratio is, i.e. the higher the frequency of a base *relative* to the frequency of its derivative, the more likely the morphological structure of the latter to be perceived, and the associated affix to be used productively.

One model which posits an explicit connection between parsing and productivity is Baayen's (1993) dual processing race model of morphological access, where the full form of a morphologically complex word such as *government* is represented as a one-unit access representation, together with units corresponding to its constituent parts (*govern* and *-ment*). Upon presentation of *government*, both the whole-word and compositional representations compete for activation, as a function of i) contextual information of previously activated units (priming), and ii) how often the units have been activated. In the absence of strong contextual effects, if the frequency of *government* is greater than that of *govern*, then the unit corresponding to the former will reach a critical selection threshold quicker, and will be used as an access representation. Accordingly, *government* is accessed as a whole. In a competition between units corresponding to *dazzle* and *dazzlement*, on the other hand, decomposition will get the upper hand, due to the greater frequency of *dazzle* over *dazzlement*. On average, for *-ment* to remain productive, words containing it must be parsed sufficiently often. In this way, the resting activation level of that affix (i.e. the level of activation at which the affix unit starts its race for reaching the selection threshold) remains high, thus conferring it a significant advantage over other possible competitors (whose starting point is lower). In this way, the model posits a strong connection between productivity and decompositional parsing in perception. High rates of decompositional access ensure the productivity of an affix. Conversely, an affix which is contained by many words accessed as a whole is unlikely to be productive.

To capture the fact that words encountered frequently have different lexical properties from words encountered relatively infrequently, all models must assume that accessing a word in the mental lexicon in some way affects the access representation of that word. However, in all these models, access representations are assumed to be about given symbolic objects, which are part and parcel of the training environment, with no or little questioning of their developmental history. Now, for *government* to be mapped onto two access units (*govern* and *-ment*), the units must be perceived and stored independently. This does not only imply a parsing stage, for the input word *government* to be split into its parts and mapped onto the corresponding access units. It also presupposes an alignment between, say, the lexical representations of *government* and *dazzlement*, for them to be perceived and recoded in terms of partially overlapping (access) representations. This point holds no matter whether one is willing to endorse a direct access model for lexical representations (with no mediation of peripheral, modality-specific access units, as proposed by Marslen-Wilson and Zhou (1999), or a mediated access model of lexical access (Forster 1976, Caramazza et al. 1988, Baayen et al. 1997). The correlation between frequency of input forms and perception (or lack of perception) of their structure, shows that it is not possible to decouple representations from the processing operations defined over representations. Access representations in the lexicon differ exactly because they are differentially processed in serial perception and storage.

## 2. Lexicon, memory and inflection

The connection between frequency and perception has been the focus of intense investigation in the literature on working memory (Gathercole and Baddeley 1989, Papagno et al. 1991), which studies the human ability to recode and retain sequences of linguistic items (e.g. letters, segments, syllables, morphemes or words). Items that are frequently sequenced together are known to be stored in long-term memory as single chunks, and accessed and executed as though they had no internal structure. This increases fluency, eases comprehension and also explains the possibility to retain longer sequences in short-term memory when familiar chunks are presented, see Cowan (2001). Even more interestingly for our present concerns, parts belonging to high-frequency chunks tend to resist being perceived as autonomous elements in their own right and being used independently. In the present paper, we set out to develop the connection between processes for short-term and long-term storage and aspects of morphological organisation in the mental lexicon.

Memory processes for serial cognition are helpful in establishing the explanatory link between the developmental course of word memory traces in the mental lexicon and their organisation and role in word perception, access and productivity. In particular, we intend to illustrate here how frequency yields the "wholeness" effect, why frequently-used words compete with members of their own lexical families (such as inflectional or derivational paradigms) and why they tend not to participate in larger series of paradigmatically homologous words (*government*, *department, argument* etc.). With this purpose in mind, we will make use of Temporal Kohonen's Self-Organising Maps (TSOMs) (Koutnik 2007, Ferro et al. 2010; 2011), which define a class of unsupervised artificial neural networks mimicking the behaviour of small aggregations of neurons in the cortical areas involved in the classification of sensory-motor data. In TSOMs, processing consists in the serial activation of specific memory nodes upon presentation of a particular time sequence of input stimuli. Through repeated exposure to such sequences, nodes get specialised for both nature and context of each stimulus, with temporal connections between consecutively-activated nodes defining the map's expectation for an incoming stimulus. TSOMs provide a general framework for putting algorithmic hypotheses of the processing-storage interaction to the empirical test of a computer implementation. Unlike classical perceptron-like neural architectures trained on back-propagation, they allow scholars to entertain a truly emergentist view of morphological competence, based on a number of realistic assumptions concerning acquisition of word structure. In the ensuing section, we focus on a few implications of this view from the perspective of lexical representations.

## 3. Emergent lexical representations

Emergentist, associative views on the (morphological) lexicon, see Bybee (1995), Bates and Goodman (1999), Burzio (2004) among others, treat word forms as primitive units and their recurrent parts as derivative abstractions over word forms. According to this perspective, full forms constitute the basis for morphological processing, with sub-lexical units resulting from the application of morphological processes to full forms. Morphology acquisition amounts to learning the relations between fully-stored word forms, which are concurrently available in the speaker's mental lexicon and jointly facilitate processing of morphologically-related forms.

In a network-based interpretation of the associative view (Bybee 1995) word forms sharing meaning components and/or phonological structure are associatively connected with one another, as a function of formal transparency, item frequency and size of morphological family (Fig. 1). In the figure, phonological transcriptions of a few Italian verb forms are assumed to be stored independently, with shared sound sequences being mutually linked through associative connections. Dashed lines represent connections between words that are only formally similar, and solid heavier lines represent connections between words that are both formally and semantically similar. Hence, different forms of the same verb (paradigmatically-related forms) and forms sharing the same inflectional ending (paradigmatically homologous forms) are connected through solid lines, with nodes corresponding to shared endings being highlighted in grey. Similar-sounding words are connected by dashed lines only.

According to Bybee, the strength of lexical connections is affected by frequency. High-frequency word forms have greater lexical autonomy, i.e. their lexical connections with other morphologically related forms are weaker. Hence, the strength of a pattern is inversely proportional to the number of times a particular sequence (a full form, a stem or an affix) instantiates the pattern. On the other hand, the strength is directly proportional to the number of different contexts where the sequence is found, i.e. to the number of outgoing connections leaving a particular node, or a particular sequence of nodes, entering the pattern (see for example, the number of connections emanating from the inflectional ending *-o* in *vengo*, *tengo* and *temo*).

**Figure 1**: A network-based model of an associative lexicon containing a few Italian verb forms. The network represents 3 forms of the verb VENIRE ‘come’ (*vengo* ‘I come’, *vieni* ‘you come’ and *vengano* ‘they come’ subjunctive), 2 forms of the verb TENERE ‘keep/hold’, and the form *temo* ‘I fear’. Verbs are provided in phonological transcription (adapted from Bybee 1995).

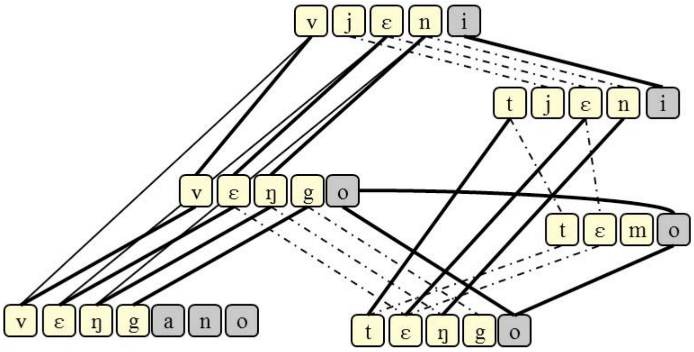
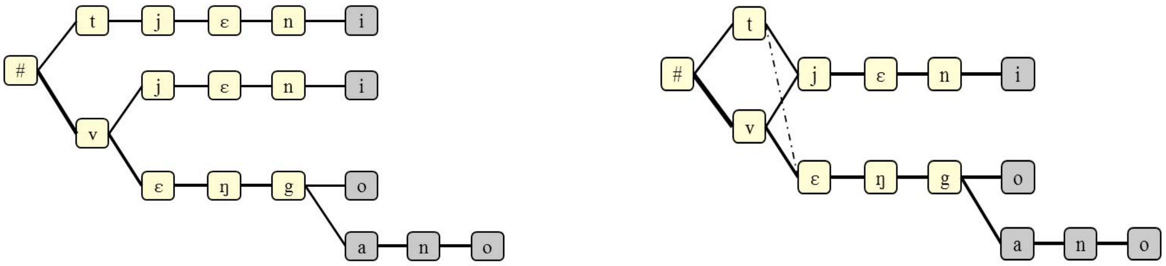


**Figure 2**: A tree-based (a) and a graph-based (b) representation of four phonologically transcribed Italian verb forms
a) b)



The tree-like representation of concurrently-stored forms in Fig. 2a exemplifies a different type of data structure. Redundant patterns are represented by shared chains of connected nodes, where nodes stand for basic representational units (e.g. letters or sounds), and directed (rightwards) arcs link two consecutively occurring units. In a probabilistic interpretation of a word-tree, the strength of each connection reflects how often the units corresponding to connected nodes are seen one after the other. Hence, a high-frequency form tends to develop a chain of strongly connected nodes. The strength of connections defines the level of entrenchment of a form in the lexicon and can be interpreted dynamically as the conditional probability with which a particular form is expected to occur, when an increasingly longer part of the word is perceived. In the tree, word forms belonging to the same paradigm (say *vengo* 'I come' and *vengano* 'they come' subjunctive) share a chain of nodes representing the common stem. The chain bifurcates upon encountering different inflection suffixes (*-o* and *-ano* respectively). In a tree, any two forms never meet again after the first bifurcation point. Traversing a tree-like representation of the lexicon, from its start-of-word symbol ('#') down to the leaf of any particular branch, simulates the process of progressively narrowing down the range of lexical entries sharing the same onset, until a point is reached where only one alternative remains available, in the spirit of cohort models of lexical access (Marslen-Wilson 1987). Fig. 2b depicts yet another data structure for the same set of forms. Whenever possible, separate branches may converge as the result of two words sharing the same tails (e.g. *vieni* 'you come' and *tieni* 'you hold'). The structure allows any node to be reached by more rightwards connections at the same time. This produces a considerable reduction in the number of nodes needed to represent a set of morphologically-related words.

In previous work (Marzi et al. 2012a; 2012b), we showed that TSOMs dynamically simulate self-organisation processes leading to the data structures of Fig. 2. To better understand the relationship between frequency, entrenchment, morphological parsability and productivity in these processes, we need to know more about the use of TSOMs as models of lexical storage and processing. It is important to emphasise at this stage that the main difference between the lexical network in Fig. 1 and the data structures of Fig. 2 lies in the way associative relations are modelled. Tree-like and graph-like structures make use of lexical connections between consecutively occurring units to model both entrenchment of individually stored forms and associative relations between concurrently stored forms. On the other hand, network models resort to two different mechanisms (lexical entrenchment and lexical association) to account for the inverse correlation between frequency and lexical productivity. A graph-like approach is more in line with recent theoretical models of emergent lexical organisation, e.g. Burzio (2004) and neuro-functional architectures of the language processor, e.g. Catani et al. (2005), which blur the distinction between storage and computation, along with the dichotomy between morphological representations and morphological rules. We will return to these points in the concluding section.

## 4. Self-organising maps for lexical storage and processing

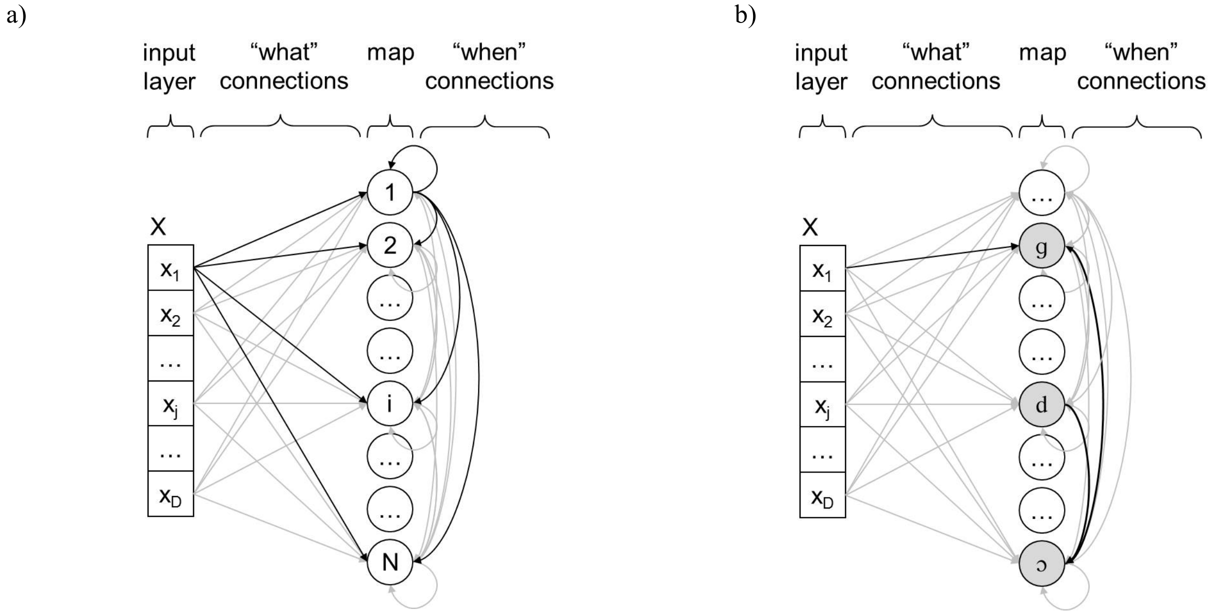
A TSOM is a grid of topologically-arranged memory nodes trained to selectively respond to classes of input stimuli occurring in specific temporal contexts (Pirrelli et al. 2011). Neighbouring nodes on a TSOM are activated by similar input stimuli, where similarity reflects the nature of the stimulus and its time-bound distribution.

Each map node receives information from an INPUT LAYER, representing the most peripheral level of input encoding (see Fig. 3a). "What" connections define the communication channel between the input layer and the map proper. On top of that, each map node communicates with any other node through pre- and post-synaptic weighted connections, referred to in Fig. 3a as "when" connections. TSOMs define an interesting class of artificial neural architectures for simulating processes of lexical organisation.

### 4.1 Recoding

At an appropriate level of abstraction, word forms consist of temporal sequences of sensory stimuli. The lexical form /'dɔg/, for example, can be input to a TSOM as a time-bound signal, made up out of three segments presented at consecutive time steps. At each step, the most highly activated node responding to the current stimulus (or *Best Matching Unit*, hereafter BMU) is selected as the "winner" and represents the map's response to the stimulus. After presentation of the form /'dɔg/, the map will have selected three different winners (BMUs), one at each time step. The resulting CHAIN of BMUs represents the way the temporal stimulus /'dɔg/ is RECODED and perceived by the map (Fig. 3b). Input recoding is thus based on the integrated pattern of node activation resulting from exposing a TSOM to an input word. At this level, the map caches recurrent processing steps through memory nodes, and it uses them over again whenever possible. Hence, recoding lies at the core of the map organisation. Stimuli that have been processed by identical or neighbouring nodes will be recoded through identical or neighbouring nodes. Patterns of node activation thus give information about how close two stimuli are processed and eventually stored by a TSOM.

**Figure 3**: a) outline architecture of a TSOM; b) an activation chain at time step 3, after the last segment /g/ of /'dɔg/ is input to the TSOM.



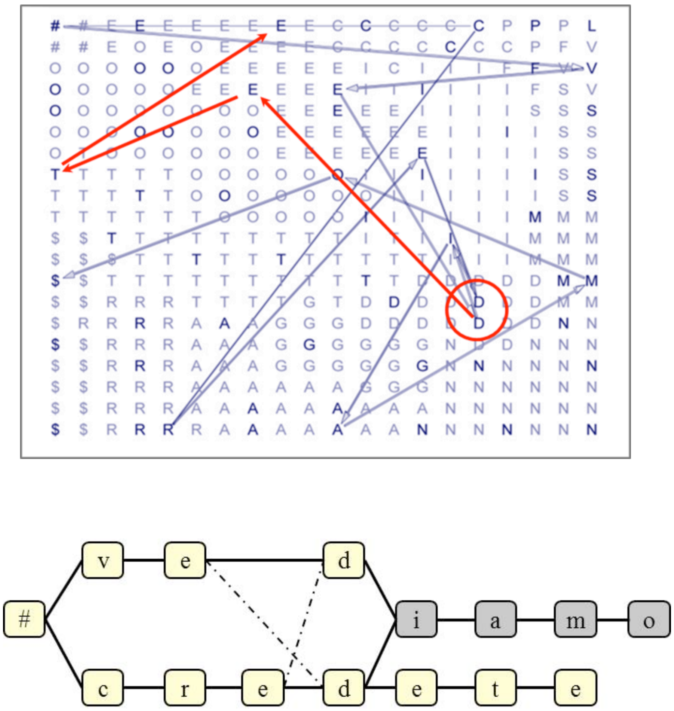
## 4.2 Training

Unlike nodes in feed-forward back-propagation neural networks, map nodes are not pre-wired to respond to specific classes of stimuli. In fact, recoding is the end result of training a TSOM on a representative set of stimuli (e.g. a sample of inflected verb forms with their frequency distribution in a reference corpus).

Training consists in showing the map one input form at a time, each sampled according to its distribution. All forms are sampled in one training epoch and the process is repeated over again for several epochs (e.g. 100). During training, nodes that respond most strongly to specific stimuli (BMUs) get increasingly attuned to the distinctive features of those stimuli. Since nodes are sensitive to both the nature of an input stimulus (e.g. the acoustic features of the sound /g/ in /'dɔg/) and its temporal context (the fact that the sound occurs in word final position, or it is preceded by a vowel), a TSOM is likely to develop many different nodes for the same stimulus, with each node being specialised to respond to a particular contextual realisation of the stimulus.

When one node is selected as the current BMU, its connections are adjusted. "What" connections become closer to current values on the input layer (Fig. 3b). At the same time, "when" connections between the current BMU and the BMU at the immediately preceding time step are strengthened. Since the overall activation level of a map node is the summation of input activation (flowing from the input layer through "what" connections) and context activation emanating from the previous BMU (through "when" connections), an adjusted BMU will be more likely to win over again when the same stimulus is presented to the map in the same context. Due to this training dynamic, selective specialisation of nodes is the natural bias of a TSOM. Other things being equal, a TSOM tends to structure the lexicon as a word-tree (Fig. 2a).

**Figure 4**: BMU activation chains for *vediamo-vedete-crediamo* on a 20×20 map (top) and their word-graph representation (bottom).



### 4.3 Redundancy in a lexical TSOM

However, things are not always equal. Concurrently stored forms may compete for the same pool of memory nodes due to the interplay of several factors: i) amount of available memory resources; ii) "wordlikeness" (i.e. amount of shared formal redundancy) of input words; iii) frequency of input words; iv) plasticity of the map. If an input word is not wordlike, i.e. if it consists of a sequence of symbols which is not typical of (many) other words in the lexicon, it is more likely to activate dedicated nodes. If a wordlike input word is shown to a TSOM only occasionally, it will fail to develop dedicated nodes. Conversely, high-frequency words have a tendency to develop diverging branches of nodes. Finally, plasticity defines the map's readiness to adjust connection weights. During training, the map loses its plasticity, so weights are adjusted less and less adaptively as training progresses.

Figure 4 shows two activation chains for the Italian verbs *crediamo* ('we believe') and *vediamo* ('we see') on a 400-node TSOM trained on Italian inflected verb forms. On the map, each node is labelled with the symbol (letter) the node responds most strongly to. Solid lines represent Hebbian connections linking consecutively activated BMUs. The two chains thus describe the path the map goes through upon recognising the two input words, one letter at a time. The paths are more distant on the roots *cred*- and *ved*-, and tend to converge topologically as soon as more letters are shared by the input forms. Eventually, the substring -*iamo* activates the same BMUs. A more symbolic representation of the two chains is offered by the directed word graph (Fig. 3, bottom), where vertices correspond to map nodes, and arcs stand for "when" connections. It can be shown (Marzi et al. 2012b) that co-activation of a pool of BMUs by morphologically-related input words i) reflects the extent to which the map perceives their formal relationship and ii) is a logical precondition to morphological generalisation.

## 5. The dynamic of word acquisition

To address issues of frequency, productivity and developmental acquisition, we ran two experiments. In the first experiment, we intended to investigate the interconnection between time of acquisition and frequency distribution of inflectional paradigms. In particular, we wanted to understand to what extent word frequency affects acquisition, and what is the role of the relative frequency distribution of forms belonging to the same paradigm in the acquisition of the whole paradigm. In a second smaller-scale experiment, we used artificial "mini-paradigms" to explore the incremental behaviour of paradigm cells with respect to issues of morphological generalisation. The experiment was intended to understand the dynamic of inter-paradigmatic COACTIVATION of paradigmatically-homologous forms as an explanatory basis for morphological productivity.

### 5.1 Experiment 1: the time-course of paradigm acquisition

The time-course of lexical acquisition is known to be affected by several factors, ranging from word length, word frequency and time of acquisition, to wordlikeness, perceptual salience and even emotional valence. From a memory-based perspective, it makes sense to focus on a few low-level, pre-theoretical and even pre-morphological factors influencing this dynamic. We suggest taking this perspective seriously for several reasons: it is open to computational and algorithmic investigation, it focuses on some founding, cognitively-grounded factors which have extensively been explored in the relevant literature, and it allows establishing a more direct connection between behavioural evidence and neuro-functional aspects of word processing. For this purpose, we first monitored the time course of acquiring a representative sample of realistically-distributed German verb forms, to establish their epoch of acquisition. We then compared this time-course with another artificially-induced time-course of the same forms, under the assumption that forms were presented to a map with a uniform frequency distribution (as opposed to real corpus-based distributions).

#### *5.1.1. Materials and methodology*

We selected from CELEX (Baayen et al. 1995) the top 50 high-frequency German verb paradigms. From each paradigm belonging to the original sample, we extracted 15 inflected forms including the full set of present indicative and praeteritum forms, the past participle, the infinitive and the present participle. All forms were encoded as strings of capitalised letters, starting with '#' and ending with '$'. Umlauted characters were encoded as lower-case digraphs (e.g. '#HoeREN$' for *hören*) and the sharp s 'ß' as 'ss' (e.g. '#HEIssEN$' for *heißen*). In both cases, pairs of lower-case letters are processed as one symbol. All letters were encoded on the map's input layer as mutually orthogonal binary vectors.

Each input word was administered to a TSOM one letter at a time, with "when" connections being reset upon presentation of '#'. We experimented with two training regimes. In one regime, input forms were administered according to a function of their frequency distribution in CELEX, to simulate more realistic conditions of input exposure. In the second regime, all forms were shown to the map the same number of times (5 each). We trained 5 different maps under each training regime for 100 epochs. Finally, we compared the behaviour of the two groups of maps (realistically-trained maps vs. uniformly-trained maps) on two tasks: word recognition and word recall.

Word recognition consists in recoding an input form as an activation chain of BMUs over the map. Errors occur when an input letter activates a BMU associated with a different letter. An input word is recognised correctly if each BMU in the activation chain is correctly associated with the current input letter. Word recall simulates the reverse process of retrieving a sequence of letters from an activation chain of BMUs. This is achieved through spreading of activation from the start-of-word node ('#') through the nodes making up the activation chain. At each time step, the map outputs the individual symbol associated with the currently most highly-activated node. The step is repeated until the node associated with the end-of-word symbol ('$') is output. Errors occur when the map misrecalls one or more symbols in the input string, by either replacing it with a different symbol or by outputting correct symbols in the wrong order. Partial recall, i.e. the correct recall of only a substring of the target word (say ‘#GEB$’ for ‘#GEBE$’), is also counted as an error. Due to the dynamic interplay between short-term processing (recognition) and long-term storage (acquisition), for a word to be recalled accurately, it must first have been recognised accurately. Hence, correct word recall tend to take place a few epochs after correct word recognition.

### *5.1.2. Results*

**Figure 5**: a) the time course of lexical acquisition of realistically distributed (red line) vs. uniformly distributed (blue line) lexical forms; results are provided for both type (solid lines) and token counts (shaded line) as a fraction of all input words (recall accuracy); b) average frequency of correctly recalled words by learning epochs; c) average length of correctly recalled words by learning epochs. Counts are averaged over the 5 instances.

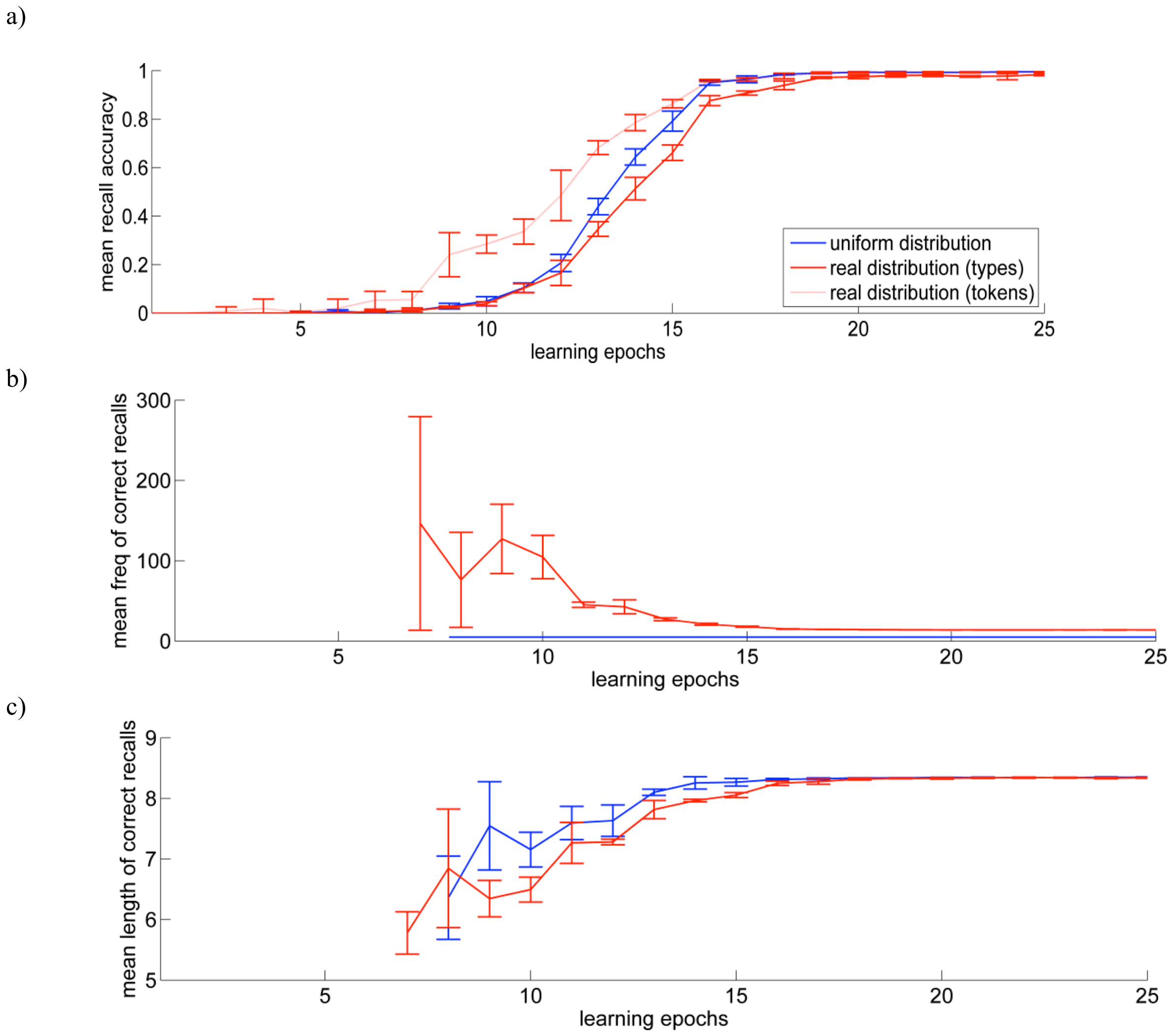


Fig. 5a plots the incremental time course of word acquisition for the two groups of TSOMs: the realistically-trained and the uniformly-trained maps. For any given word form, we define its time of acquisition by a TSOM as the first epoch starting from which the form is RECALLED correctly. Unlike word recognition, which mostly depends on the current input stimulus, word recall entirely depends on internal recoding, and requires that fine-grained information about the nature and timing of each symbol is stored in the internal state of the map. The plot shows how many words are acquired at each epoch, as a fraction of all input words. Hence, unity means that all input words were acquired (by being recalled correctly). Counts are averaged over the 5 instances of each group, with standard deviation represented by whiskers.

Results are shown by both TYPE and TOKEN frequency. By counting types, we consider the number of different forms[1] that are accurately recalled at each epoch, and divide the number by the total number of different forms used for training (750). By counting tokens, we consider the number of times all forms that are recalled at each epoch appear in the training set (as a function of their frequency distribution in CELEX), divided by the overall number of times the map is exposed to all words (10,286). For the realistically-trained map group, we show results by both type and token frequency counts. Clearly, for the uniformly-trained map group, type and token results coincide.

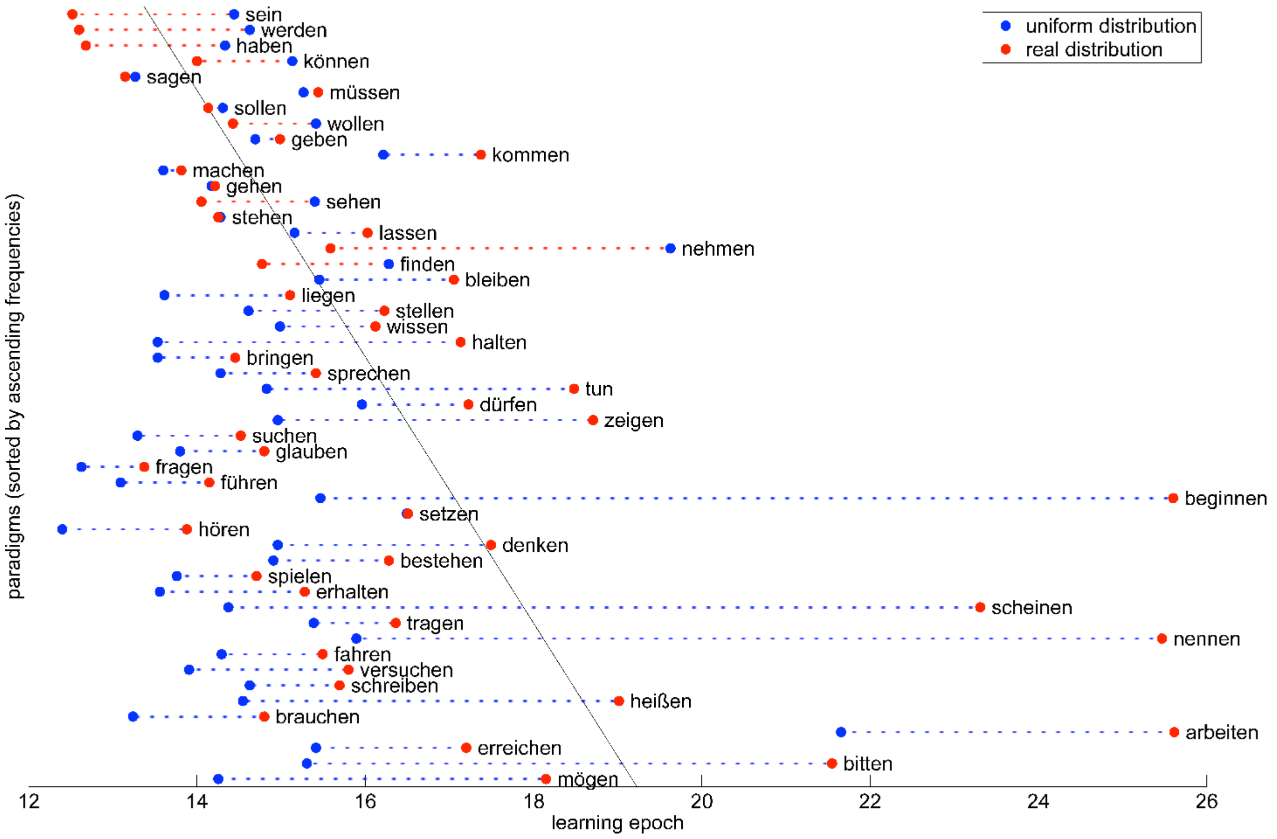
Note that the realistically-trained group acquires word types more slowly than a uniformly-trained group does, particularly in the 10-15 epoch range, suggesting that there is a statistically significant advantage ($p < 0.005$ at epoch 30) in having all forms presented an equal number of times during training. This reflects a learning bias of TSOMs. The training algorithm minimises the most frequently repeated mismatches between an input stimulus and its internal recoding. Thus, a map trained on realistically-distributed data tends to acquire first those forms that happen to be encountered more often, while "neglecting" less frequent words, as shown by the mean frequency of correctly recalled forms at each epoch (Fig. 5b). This also negatively correlates with word length (Fig. 5c), with longer words being learned comparatively later than shorter words. Due to the inverse correlation between word frequency and word length in the lexicon (more frequent words are generally shorter than less frequent words), maps trained on real distributions tend to consistently acquire shorter words across learning epochs than uniformly-trained maps do.

In TSOMs, words are not memorised as isolated wholes and word surface frequency is not the only frequency factor affecting acquisition and lexical organisation. A TSOM is also demonstrably sensitive to the frequency of sublexical patterns, i.e. strings that are shared by more words. As a general trend, high frequency words are learned first, and low-frequency words that share some recurrent patterns with high-frequency words are learned more quickly than isolated words. The role of formal redundancy in lexical acquisition is clearer when word frequency distributions tend to be uniform or strongly balanced. When most input words are presented equally often, the map will tend to memorise frequently-occurring sublexical patterns first. We can say that a map trained on a uniform distribution of forms puts a premium on general patterns of formal redundancy, thus memorising words by regularity rather than by frequency. Although the notions of regularity and frequency are often correlated in the lexicon (e.g. regular patterns are systematically repeated both intra- and inter-paradigmatically in regular paradigms), there are cases, such as highly frequent but paradigmatically isolated inflected forms, where this correlation is reversed. As a result, irregular paradigms (i.e. paradigms containing many alternating stems) would be harder to learn if they were distributed uniformly. This is clearly shown when we move from the time course of word acquisition to the time course of paradigm acquisition.

For each paradigm, we define its time of acquisition by a TSOM as the mean epoch at which all forms of the paradigm are RECALLED correctly. Mean acquisition epoch of a paradigm thus provides an estimate of the average time it takes for all forms of the paradigm to be recoded in a time-sensitive way, and be recalled accurately from their corresponding activation chains. In Fig. 6, we plotted the acquisition time of each German verb paradigm over both training conditions, CELEX-based distribution (red dots) and uniform distribution (blue dots). On the vertical axis, paradigms are ordered by ascending frequency values. The black line interpolates red dots, regressing frequency ranks on acquisition epochs in the realistic training condition, and showing an inverse correlation between paradigm frequency and time of acquisition. A brief inspection of the graph confirms what we observed for individual words. In the vast majority of cases, paradigms are acquired more rapidly when they are presented with a uniform frequency. A closer look reveals that only a few high-frequency highly irregular paradigms (e.g. SEIN, WERDEN), whose alternating stems are extensively attested, show an advantage in the skewed training condition. We shall return to this point in the general discussion.

[1] For our present purposes, homographic forms instantiating different paradigm cells (e.g. *geben*) are counted as distinct word types.

**Figure 6**: Mean acquisition time of paradigms ranked by ascending frequencies in the CELEX-based training condition. Blue dots represent averaged acquisition epochs in the uniform training condition, and red dots averaged acquisition epochs in the CELEX-based training condition. The black line interpolates red dots only, showing an inverse correlation between paradigm-frequency and learning epoch.



## 5.2 Experiment 2: the role of frequency in word parsability and productivity

Experiment 1 shows some non trivial aspects of the learning dynamic of a complex inflectional system. We observed that high-frequency words are learned first, as well as high-frequency paradigms. However, the vast majority of paradigms (and the totality of regular paradigms), are learned more efficiently under a uniformly-distributed training regime. We suggest that this behaviour is due to uniformly-trained TSOMs being able to organise recoded words in a deeply interconnected network of associations, where more nodes are shared and connections are less deeply entrenched.

A more entropic map should be in a better position to generalise to unknown words. To understand more about this dynamic, we ran a second experiment using artificial mini-paradigms, small collections of abstract letter strings whereby we can control the considerable complexity of learning a real paradigm-based inflectional system by making simplifying assumptions on the nature of the training set.

### *5.2.1. Materials and methodology*

A 64 node map was trained on the same artificial data set, with four different training regimes, as illustrated in Table 1 below.

**Table 1**: Frequency distributions of two artificial mini-paradigms in four different training conditions

| paradigm id | items | Frequency | | | |
|---|---|---|---|---|---|
| | | condition 1 | condition 2 | condition 3 | condition 4 |
| 1 | #BDY$ | 5 | 100 | 5 | 5 |
| 1 | #BDZ$ | 5 | 100 | 5 | 5 |
| 1 | #BDX$ | 0 | 0 | 0 | 0 |
| 2 | #ADZ$ | 100 | 100 | 5 | 5 |
| 2 | #ADX$ | 100 | 100 | 5 | 100 |
| 2 | #ADY$ | 0 | 0 | 0 | 0 |

The training set consists of two artificial mini-paradigms: {#BDY$, #BDZ$, #BDX$} and {#ADZ$, #ADX$, #ADY$), each including two attested forms and one unknown form (training frequency = 0). There are two uniformly-distributed training sets (conditions 2 and 3), and two skewed ones (conditions 1 and 4). In condition 1, forms are distributed evenly within each paradigm, but with different frequencies across paradigms. In condition 4, there is a difference in the frequency distribution of the two paradigms (totalling respectively 10 and 105 tokens), and a difference in the distribution of paradigm members (5 and 100) for one of the two paradigms.

A 64 node map was trained for 100 epochs on each regime 10 times, with the same input protocol of Experiment 1. Eventually we examined the behaviour of 40 trained TSOMs. Each map was then tested on word recall for both attested (training) and non-attested (test) word forms. Results are aggregated and averaged by the four conditions of training.

### *5.2.2. Results*

On average, uniformly-distributed paradigms (training conditions 2 and 3) reach a stable acquisition state at epoch 7.3 and generalise to unknown forms at epoch 7.5. Their per-form dynamic through learning epochs is remarkably similar, in spite of their considerable difference in terms of absolute frequencies. Although, in training condition 2, forms are presented 100 times each, compared with 5 times each for training condition 3, their averaged learning curve is almost indistinguishable. This observation confirms the hypothesis that absolute frequencies are not as important as relative frequencies are, and provides further support to the idea that learners are more sensitive to entropy-based effects than to mere frequency distributions Moscoso et al. (2004).

Unevenly-distributed paradigms (training conditions 1 and 4) get to a stable acquisition state between epoch 7 and epoch 7.3 on average, but generalise statistically significantly later: at epoch 8. Although data are fairly preliminary, due to the small scale of this experiment, they confirm that frequency is certainly helpful for entrenchment, but it is less conducive to generalisation. Competition is the most powerful determinant of this frequency-based dynamic, as shown by the rate of acquisition of individual forms in low-entropy vs. high-entropy paradigms. The form #ADX$ which is presented 100 times in both training conditions 1 and 4, is learned more quickly and more consistently when it is in the company of a low-frequency (rather than high-frequency) paradigm member (#ADZ$). But although #ADX$ is acquired more rapidly in a less competitive context, the overall paradigm of #ADX$ is acquired more slowly than in the other, more competitive condition. This is because the rapid consolidation of a deeply-entrenched chain eventually hinders consolidation of other lower-frequency forms of the same paradigm.

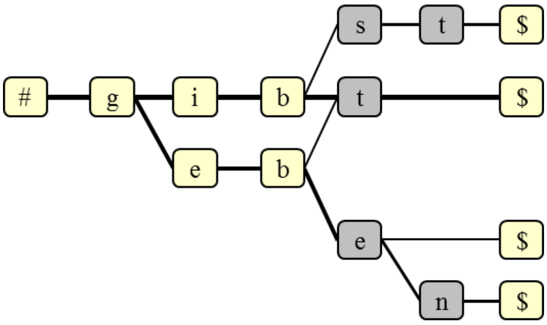
Note that the crucial contrast condition in frequency distribution is provided WITHIN each paradigm, rather than ACROSS paradigms, suggesting that there is competition between members of the same paradigm. This is confirmed when we look at the rate of acquisition of #ADX$, which also affects the developmental dynamic of generalisation to #BDX$, a test form belonging to a different paradigm. #BDX$ can in fact take advantage of the early consolidation of #ADX$ in long-term memory, exploiting the final part of its chain (-DX$). On the other hand, it is more difficult for a paradigm dominated by a high-frequency form, to generalise to novel forms.

## 6. General discussion

Frequency is the most powerful determinant of storage: highest-frequency forms are the first ones to be learned. However, since all words are concurrently engaged in the "acquisition race", what matters is the ranking of competitors by frequency, rather than their absolute frequencies. In the end, the relative distribution of frequencies shapes both topological organisation and time course of acquisition.

Frequency reverberates on all levels of lexical organisation. There are frequency effects for individual words, as well as frequency effects above the word level (e.g. inflectional paradigms), and below the word level (sublexical patterns). There is a hierarchy of frequency effects, which has far reaching consequences on the time course of lexical acquisition. It is important to appreciate that what favours entrenchment on one level can be a detrimental factor on another level. This is due to the effects of frequency on memory entrenchment, and to the effects of entrenchment on the organisation of stored words.

**Figure 7**: A graph-based representation of the present indicative forms of German GEBEN ('give'), Thicker arcs represent stronger associative connections and can be interpreted probabilistically as expectations on incoming stimuli.



Generally speaking, entrenchment is known to favour individual access and holistic perception, and disfavour coactivation (spreading of activation to other neighbouring forms) and perception of internal structure. Our simulations provide a possible explanatory mechanism of this effect and an account of its impact on paradigm acquisition. The deeply-entrenched activation chain associated with a high-frequency form will tend to attract other forms of the same paradigm (to the extent they share the same stem). This is detrimental for acquisition of other forms of the same paradigm, either by exposure o by generalisation. Fig. 7 illustrates the graph-based organisation of a fraction of the paradigm of German GEBEN on a map trained on CELEX-based frequencies. In CELEX, the form *gibt* (3rd person singular) is considerably more frequent than any other member of its own paradigm. The immediate effect of the early entrenchment of *gibt* is to strongly attract other forms of GEBEN, which "parasitically" exploit its activation chain. However, this makes acquisition of these forms more difficult in the end, since the map develops a strong expectation for one particular ending (present indicative, 3rd person singular) and perceives other endings as far less likely to occur. In both word recognition and recall, deeply-entrenched chains are more difficult to be abandoned once they are activated.

In low-entropy inflectional paradigm, few forms recur highly frequently, followed by some medium-frequency forms and a long tail of low-frequency words. Accordingly, high-frequency forms are learned at early epochs, followed by lower frequency forms that can benefit from this early acquisition by exploiting the activation chains developed by their predecessors. But the role of path-breaking high-frequency forms may have a cost for learning. At the beginning, the map devotes most resources (in terms of number of nodes and connections) to memorise them. After this initial stage, it takes longer for the map to restore a balance and make room for other low-frequency items. The main effect of this dynamic is that unevenly distributed word paradigms are learned more slowly, since they require a redistribution of memory resources to make room for lower-frequency members. The more low-frequency items are there to be learned, the longer it will take for resource reallocation to restore the balance. This general trend is confirmed by the time course of the 50 (partial) German paradigms in Experiment 1, showing a clear advantage in the learning race for uniformly-distributed paradigms.

Morphological (ir)regularity interacts with this dynamic in interesting ways. In German verb inflection, morphologically strong verbs can exhibit extensive stem alternation (e.g. *geben*, *gibst*, *gab*, or *finden, fand, gefunden*). If frequencies are uniformly distributed, then stem alternants are learned more slowly, with a prolonged period of over-generalisation. On the other hand, irregular alternants are usually preserved by their relatively high token frequency. Hence, there may be an advantage in learning unevenly-distributed paradigms if they happen to contain high-frequency alternating stems, as shown by the comparatively small group of high-frequency paradigms in the top left corner of the regression plot of Fig. 6.

On the other hand, regular paradigms always benefit from uniform distributions. Since all forms of a regular paradigm share the same default stem, all of them will contribute to a rapid consolidation of the corresponding activation chains. Under the hypothesis of a uniform distribution of these forms, the map will develop these chains concurrently, in a smooth, incremental way. Hence, such a high-entropy map is more conducive to lexical acquisition. This is again confirmed by Experiment 1, showing that all regularly-inflected verb paradigms are learned earlier if they are presented with uniform distributions.

A balanced competition for memory resources slows down the learning rate of individual words, but eventually ensures a more effective allocation of resources and a smoother convergence towards global lexical organisation. This general trend interacts with intra-paradigmatic redundancy and (ir)regularity. Due to the frequency-by-irregularity interaction, skewed distributions may in fact favour acquisition of those paradigms exhibiting extensive stem alternation, by giving stem alternants the frequency boost necessary to offset the prevalence of default stems.

What about generalisation? How do frequency distributions affect the propensity of a map to acquire non-attested forms by a generalisation step? Experiment 2 on artificial micro-paradigms was intended to preliminarily address these questions. Once more, we observed that uniform distributions tend to favour generalisation to unknown word forms. We suggest that the reason for this behaviour is lexical competition. A stored stem appearing in a high-frequency form (e.g. *gib-* in *gibt*) builds up a strong expectation for one specific ending to follow. Hence, it is considerably more difficult for such a stem to accept a novel ending than for another stem in a low-frequency form. All in all, what makes paradigm generalisation difficult is the same mechanism that slows down acquisition of input forms in the first place: frequency-based entrenchment. Entrenchment triggers a predictive behaviour of the map, based on probabilistic expectations. If there is a strong expectation for a possible ending, other alternatives are disfavoured. This inhibits the propensity of the map to acquire both low-frequency endings and completely novel endings from other paradigms.

To sum up, the most favourable internal state for learning is an unbiased one (highest entropy of competing alternatives), as this state most heavily depends on the current stimulus, rather than on the map's internal expectations. In order words, a balanced state between competing alternative connections is most faithful to the incoming signal, and, eventually, the most open one to generalisation. It should be appreciated that what represents an advantage at the level of single word learning, turns out to be a disadvantage at the level of paradigm learning. The paradigm appears to define the morphological domain where frequency effects are perceived at the level of lexical organisation. It is not clear, at the present stage of our investigation, if other morphological families, such as the series of forms sharing the same paradigm cell (e.g. all present indicative first person singulars) exhibit the same effects. We have reasons to believe that they do, although the effect is partially obscured by the path-breaking role of deeply-entrenched chains in the acquisition of other forms in the same morphological series. More experimental evidence will be needed to address these points.

## 7. Conclusion

Computer simulations of lexical processing and storage provide a methodological middle ground for testing models of word acquisition. Tracking the time course of paradigm acquisition at a fine level of detail is notoriously hard, due to i) the difficulty of monitoring levels of metalinguistic awareness in a developmental perspective and ii) the existing gap between psycho-cognitive hypotheses of lexical architecture and recent acquisitions of the neuro-functional basis of the perisylvian language network.

Through careful data analysis of the computational behaviour of TSOMs, we gained specific insights into these issues and suggested possible modelling and explanatory mechanisms together with their neuro-functional correlates. Simulations of the incremental acquisition of German verb paradigms in different training conditions support the hypothesis that perception of structure (parsability) and morphological productivity strongly correlate in the inflectional lexicon. In particular, by monitoring longitudinal progress in storage and generalisation of differently distributed inflectional paradigms, we showed that: i) high-frequency forms are stored and accessed significantly earlier than low-frequency forms; ii) deeply entrenched forms tend to block usage of other forms in the same paradigm.

Any cognitively-motivated hypothesis of lexical architecture must assume that accessing a word leaves its traces in the lexicon. This is a logically necessary step to take, if one wants to model the fact that high-frequency words have different characteristics from low-frequency words. Successfully accessing an item must hence have two consequences: i) modify the item's representation and ii) increase the probability that the item will be successfully processed in the future. Existing models capture this process in two fundamentally different ways: a) by raising the resting activation level of the relevant lexical entry (see, e.g. Norris et al. 2000, McClelland and Elman 1986); or b) by assuming that processing a word involves adding a new exemplar to the appropriate exemplar repertoire (e.g. Johnson 1997a, 1997b; Daelemans and van den Bosch 2005).

All these models assume that accessed representations already exist, thus making a fundamental distinction between representations on the one hand, and processes applying to representations on the other hand. We propose to refer to them as “distinctive” models, to mean that they apportion issues of representation and issues of processing to logically distinct modelling mechanisms. According to a “distinctive approach”, the lexicon is a box, words are its content, storing words amounts to placing content into the box, accessing words corresponds to the reverse process of getting words out of the box. Last but not least, accessing a word representation implies that something changes in the accessed content. However, the approach appears to neglect that lexical representations are acquired dynamically, and little is understood in modelling lexical storage and access if one is not in a position to explain how representations come into existence in the first place.

The present contribution offers a computationally explanatory basis to address this fundamental issue. From an acquisitional standpoint, words do not define an independently-given content, but are treated like input stimuli causing a particular change in the activation state of the lexicon. Conceptually, the activation state represents what is perceived by a memory map after input exposure and it is not to be confounded with the stimulus itself. In addition, such an internal state presents a short-term and a long-term dynamic. In the short-term, it consists in a chain of consecutively responding BMUs, whose level of activation decays as soon as the map is exposed to novel inputs[2]. In the long-term, BMUs forming the current chain are modified incrementally through an adaptive caching process, for them to be more likely to be activated when the same input is presented to the map. Clearly, no distinction is made here between representations and processing. BMUs are both representational units, i.e. the specialised, long-term activation patterns

[2] In fact, we can assume that the level of activation of a BMU at time *t* starts diminishing as soon as another input unit is presented to the map. This is one of the most common assumptions made in the memory literature to explain so-called recency effects.

indexing individual input stimuli, and processing units, dynamically responding to particular classes of stimuli. For this reason, we suggest calling this type of approach “integrative” as it deals with lexical representation and lexical processing as the same process on two different time scales[3].

This is in line with what we know about the neuro-architectural basis of the human language processor, supporting an integrative account of lexical processing/acquisition as the complex result of general-purpose operations on word stimuli: e.g. working memory, long-term storage, sensory-motor mapping, rehearsal, unit integration, unit analysis, executive control, time-series processing (Catani et al. 2005, Shalom and Poeppel 2008, Friederici 2012). Our investigation credits the proposed computational framework with psycholinguistic plausibility, and grounds parsability-based models of morphological productivity on a specific, explicit proposal of lexical architecture. This provides an explanatory basis for both psycholinguistic and linguistic accounts of morphological structure, and offers an intermediate framework for scientific inquiry bridging the gap between linguistic units and functional units in neurosciences. Finally, it makes the interesting suggestion that principles of morpheme-based organisation of the mental lexicon are compatible with a learning strategy requiring memorisation of full forms.

## References

Baayen, R.H. (1993) On frequency, transparency and productivity. In Booij, G. & J. van Marle (Eds.), *Yearbook of Morphology 1999*. Dordrecht: Kluwer, 181-208.

Baayen, R.H., T. Dijkstra, & R. Schreuder (1997) Singulars and Plurals in Dutch: Evidence for a Parallel Dual-Route Model. *Journal of Memory and Language* 37(1), 94-117.

Baayen, H., R. Piepenbrock & L. Gulikers (1995) *The CELEX Lexical Database* (CD-ROM). Philadelphia: Linguistic Data Consortium.

Bates, E. & J. C. Goodman (1999) On the emergence of grammar from the lexicon. In B. MacWhinney (Ed.), *Emergence of Language*. Hillsdale, NJ: Earlbaum. 29-79.

Burzio, L. (2004) Paradigmatic and syntagmatic relations in Italian verbal inflection. In J. Auger, J.C. Clements & B. Vance (Eds.), *Contemporary Approaches to Romance Linguistics*. Amsterdam/Philadelphia: John Benjamins, 17-44.

Bybee, J. (1995) Regular Morphology and the Lexicon. *Journal of Verbal Learning and Verbal Behavior* 10, 5. 425-455.

Caramazza A, A. Laudanna & C. Romani (1988) Lexical access and inflectional morphology. *Cognition* 28(3), 297-332.

Catani, M., D. K. Jones & D.H. ffytche (2005) Perisylvian language networks of the human brain. *Annals of Neurology* 57, 8-16.

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

Daelemans, W. & A. van den Bosch (2005) *Memory-Based Language Processing*. Cambridge: Cambridge University Press.

Elman, J. L. (2004) An alternative view of the mental lexicon. *Trends in Cognitive Sciences* 8(7), 302-306.

Ferro, M., D. Ognibene, G. Pezzulo & V. Pirrelli (2010) Reading as active sensing: a computational model of gaze planning in word recognition. *Frontiers in Neurorobotics* 4(6). DOI:10.3389.

Ferro, M., C. Marzi & V. Pirrelli (2011) A Self-Organizing Model of Word Storage and Processing: Implications for Morphology Learning. *Lingue e Linguaggio* X(2), 209–226.

Forster, K. I. (1976) Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New approaches to language mechanisms*, Amsterdam: North-Holland. 257-287.

Friederici, A. D. (2012) The cortical language circuit: from auditory perception to sentence comprehension. *Trends in Cognitive Sciences* 16(5), 262-268.

Gathercole, S. E. & A. D. Baddeley (1989) Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language* 28, 200-213.

Hay, J. (2001) Lexical frequency in morphology: is everything relative? *Linguistics* 39, 1041-1070.

Hay, J. & H. R. Baayen (2002) Parsing and productivity. In Booij, G. & J. van Marle (Eds.), *Yearbook of Morphology 2001,* Kluwer Academic Publishers. 203-235.

Hay, J. & I. Plag (2004) What Constrains Possible Suffix Combinations? On The Interaction of Grammatical and Processing Restrictions in Derivational Morphology. *Natural Language & Linguistic Theory* 22, 565–596.

Johnson, K. (1997a) The auditory/perceptual basis for speech segmentation. In K. Ainsworth-Darnell & M. D’Imperio (Eds.), *Ohio State University Working Papers in Linguistics: Papers from the Linguistics Laboratory 50*, Columbus: Ohio State University, 101-113.

[3] Similar “integrative” assumptions, however cast into a different computational framework, are made by Elman (2004).

Johnson, K. (1997b) Speech perception without speaker normalization. In Johnson K. & J. Mullenix (Eds.), *Talker Variability in Speech Processing*. San Diego: Academic Press, 145-165

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

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

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

Marslen-Wilson, W.D. & X. Zhou (1999) Abstractness, allomorphy, and lexical architecture. *Language and Cognitive Processes* 14, 321-352.

Marzi, C., M. Ferro & V. Pirrelli (2012a) Prediction and Generalisation in Word Processing and Storage. In A. Ralli, G. Booij, S. Scalise & A. Karasimos (Eds.), *On-line Proceedings of the 8th Mediterranean Morphology Meeting,* 113-130.

Marzi, C., M. Ferro & V. Pirrelli. (2012b) Word alignment and paradigm induction. *Lingue e Linguaggio* XI(2), 251-274.

Moscoso del Prado Martín, F., A. Kostić, & H. Baayen (2004) Putting the bits together: an information theoretical perspective on morphological processing. *Cognition* 94, 1-18.

Norris, D., J.M. McQueen & A. Cutler (2000) Merging information in speech recognition: feedback is never necessary. *Behavioral and Brain Sciences* 23(3), 299-325.

Papagno, C., T. Valentine & A. Baddeley (1991) Phonological short-term memory and foreign-language learning. *Journal of Memory and Language* 30, 331-347.

Pirrelli, V., M. Ferro & B. Calderone (2011) Learning paradigms in time and space. Computational evidence from Romance languages. In M. Maiden, J.C. Smith, M. Goldbach & M.O. Hinzelin (Eds.), *Morphological Autonomy: Perspectives from Romance Inflectional Morphology*, Oxford: Oxford University Press, 135-157.

Shalom, D.B. & D. Poeppel (2008) Functional Anatomic Models of Language: Assembling the Pieces*. The Nueroscientist* 14, 119-127.