Co-activation and competition effects in lexical storage and processing

Vito Pirrelli
Institute for Computational Linguistics, CNR Pisa, Italy

Abstract

According to traditional wisdom in Linguistics, morphologically simple words reside in the mental lexicon, a kind of brain dictionary that contains unpredictable mappings between lexical features. Here I illustrate some of the defining features of an alternative view of the language architecture, where computation and storage are just the short-term and long-term dynamics of the same underlying process. Empirical results of a computational model of this view are reported and general implications for a theory of the lexicon are discussed.

Keywords: mental lexicon, morphology, human language processing, artificial neural networks, lexical self-organisation, language acquisition, language architecture.

1. Introduction

According to traditional wisdom in Linguistics, morphologically simple words reside in the mental lexicon, a kind of brain dictionary that contains mappings between lexical features such as sound, meaning, part-of-speech, morpho-syntactic information and the like. These mappings form the atomic, unpredictable, context-invariant building blocks of language, assembled in context by rules to form larger linguistic units, from morphologically complex words, to phrases and sentences.

At the roots of this classical conception lies the assumption that the mental lexicon is an enumerative list of unpredictable items, and that lexical knowledge interacts with processing rules in a one-way fashion, providing information that is fundamentally independent of further levels of rule-based manipulation. The lexical building blocks must be there in store before the construction process begins to set in. In turn, rules must be separate from lexical entries, in so far as their working principles do not reflect the way lexical information is stored in the mind. In this connection, matters of content (e.g. “how much information is a lexical entry expected to contain?”) are assumed to be independent of matters of form (“what representational mechanism is needed to encode lexical information?”). Accordingly, in classical language architectures (from Bloomfield 1933 to Pinker & Ullman 2002, going through as diverse theoretical frameworks as Miller & Chomsky 1963, Di Sciullo & Williams 1987, Pinker & Prince 1988, Anderson 1992, Halle & Marantz 1993, Prasada & Pinker 1993, Beard 1995, Pustejovsky 1996, Levelt et al. 1999, Stump 2001 and Clahsen 2006 to mention only a few), rules and words may interact on the content level only, but they have little to share at the level of computation. In particular, the lexicon has to do with storage, and rules with computation. In turn, storage and computation are embodied by, and localized in fundamentally distinct components, as is the case with hard disks and CPUs in commercial desktop computers. Here follows a (possibly incomplete) list of the criterial features defining such a traditional “modular” view of the lexicon:
(i) **enumerative**: the list of lexical items is countable
(ii) **serial**: items are accessed individually, one after the other
(iii) **static**: lexical information is stable; information stored in one item does not change as a function of a) how many times the information is accessed, and b) information accessed in another item
(iv) **processing-independent**: the form of lexical representations is fundamentally independent of processing principles
(v) **redundancy-free**: the lexicon contains atomic, unpredictable information only, irreducible to general principles
(vi) **word-based**: the lexicon contains word-size units or units smaller than words (e.g. stems and affixes)
(vii) **one-way interaction**: lexical content affects processing, but it is not affected by processing

Although there is arguably no classical model of the lexicon that complies with the full list of defining principles above, it is nonetheless fair to say, I believe, that a significant majority of the points in the list above appears to characterise the linguistic “vulgate” of the lexicon, and its possible mental correlates, sufficiently accurately.

Over the last years, however, mounting evidence has been found that this traditional picture is, at best, an oversimplifying abstraction. Two relatively independent lines of scientific inquiry led to a new picture, blurring the divide between lexicon and grammar, storage and computation. The first line revolves around linguistic evidence from human language processing and language acquisition, to suggest that a lot of lexical information bears on processing, and that no realistic theory of the lexicon, separating linguistic knowledge from world knowledge, can account for the totality and pervasiveness of these effects. The second line of research focuses on neuro-biologically inspired computer models of language architectures. They show that the mental lexicon is part and parcel of the processing engine of the human language competence, and that any subdivision of labour between a static, enumerative list of lexical items and a dynamic rule engine (a cornerstone of very influential models of word processing, such as Pinker and Ullman’s Declarative/Procedural model) is in fact hardly tenable. Novel processing architectures call for a radically new conception of the lexicon-grammar interaction. In particular, we will consider here the consequences of this new conception on issues of word representation and processing. Before we get into that, section 2 offers a short overview of the relevant literature on the lexicon vs. grammar interaction.

2. **Blurring the lexicon-grammar divide**

2.1 **Linguistic and Psycholinguistic Evidence**

A primary source of evidence calling into question the idea of a sharp separation between words and rules involves investigations into human sentence processing. The evidence suggests that lexical representations must be quite rich, and affect the interpretation of grammatical structure all the way through (Altmann 1998; MacDonald 1997; Tanenhaus et al. 1995). To illustrate, Pustejovsky’s (1996) generative lexicon can
explain the most likely interpretation of a compound like BREAD KNIFE as a ‘knife for cutting bread’ in terms of the telic predicate CUT, which is assumed to be part of the world knowledge encoded in the lexical entry for KNIFE, as a tool for cutting. It is such a rich repertoire of semantic and world knowledge information at the level of individual lexical entries, one can argue, that explains the speaker’s ability to interpret some word combinations, like Noun-Noun compounds, where the relation linking two constituents is not overtly expressed (Lieber 2004; Pirrelli & Guevara 2012). In a similar vein, Ferretti and colleagues (2001) show that when the verb arrest is presented with no context, it facilitates naming of both cop (a typical arrestor) and crook (a typical arrestee). However, primes of the form She arrested the… facilitate naming crook, but not cop. Conversely, primes of the form She was arrested by … facilitate naming cop, and not crook. The three results, combined together, demonstrate that words, however powerful they can be when used as context-evoking cues, are only one such cues, and that richer lexical contexts have a considerable potential for evoking common sense knowledge and typical situations.

The problem with any attempt to explain lexical effects on sentence processing with a larger and richer lexicon is that it soon becomes difficult to draw a principled line between what should be encoded in the lexicon, and what should not. How specifically linguistic and context-free is lexical information? Alternatively, to what extent is it overlaid with information concerning aspects of the world, contingencies of the particular action described by a verb, or its typical role fillers, or even the aspectual time frame through which the action is expected to unfold? After overviewing the empirical debate on the last decades of evidence on mechanisms for sentence processing, Elman (2009) arrives at the somewhat paradoxical conclusion that the lexicon does not exist, as it cannot possibly be separated from other types of context-based and world-knowledge information.

The second source of evidence comes from the domain of developmental psychology, where words are understood as the foundational elements of language acquisition, from which early grammar rules emerge epiphenomenally. Both words and rules ultimately enforce form-meaning mapping relationships, and they should accordingly be cast into a unified representational format (Tomasello 2000; Goldberg 2003; Jackendoff 2007). Ultimately, the main difference between lexical patterns and general rules is that the latter are a schematised version of the former, with lexically instantiated slots being replaced by variables. For example, the rule governing the argument pattern of a transitive verb can be derived from the lexical argument pattern associated with the verb EAT by replacing some idiosyncratic bits of information (e.g. phonological form of the base, meaning, and argument selectional restrictions) with variables.

However, if rule-like patterns emerge from lexical patterns by abstracting away from word-specific idiosyncrasies, and if words are ontogenetically prior to abstractions, the problem arises of how words are acquired in the first place. Children are not normally exposed to individual words, but only to words in context, i.e. lexical forms that are already embedded in complex linguistic structures. This developmental account thus appears to make a somewhat paradoxical suggestion. Rule-like patterns emerge from words, but, at the same time, words are learned in the context of larger constructions, thus requiring the preliminary acquisition of patterns larger than words. The paradox revolves around the minimal size of the linguistic units that are acquired and stored
ontogenetically, for more complex units to possibly be understood through on-line combination/composition of already acquired units.

A way out of this paradox is to assume that children do not store (only) words but (also) lexical constructions (Bates & Goodman 1997; Tomasello 2000; Goldberg 2003). Lexical constructions, rather than words in a strict sense, are the foundational elements from which early grammar arises epiphenomenally. This approach is compatible with Jackendoff’s suggestion (2002) that the lexicon is not just a long-term repository of words, but contains much more and much less than words. For example, it can contain idiomatic expressions like KICK THE BUCKET, whose meaning ‘die’ cannot be derived compositionally as a function of the meanings of its individual constituents and their syntactic relations. On the other hand, it may contain units such as bare stems or affixes that are smaller than words, and are used to construct most grammatical words online, without having to store words in long-term memory.

Jackendoff comes to the conclusion that most of what have previously been called rules of language turn out to be lexical items too (2002: 154). Following his argument, one can argue that classical word entries, as well as grammatical rules, are the outcome of two different processes of schematisation: one generalising across all forms belonging to the same word paradigm, and the other one generalising over different paradigms of words which are embedded in the same syntactic patterns. Clearly, there is a substantive difference between this conception of rule learning and the traditional one. According to traditional wisdom, the difference between a regular and an idiosyncratic pattern is a matter of binary choice: the former contains at least one variable, while the latter exhibits no variable slots at all. In Pinker and Ullman’s Declarative/Procedural model (Pinker & Ullman 2002), for example, either a word form is constructed on line on the basis of atomic units that are stored in the lexicon, or it is stored as a full exception in the lexicon. No intermediate choice is available. In fact, what we observe from careful analysis of developmental data is the gradual emergence of constructions appearing in usage one verb at a time (Tomasello 2000; Wilson 2003). Similarly, some regularly-inflected forms are perceived as more regular than other equally regularly-inflected forms (Albright 2002). In addition, irregularly inflected forms too exhibit graded levels of predictability, as shown by the ring/rang/rung pattern in English.

Once more, the empirical question that arises in this connection concerns the amount of information that more or less schematised lexical patterns are assumed to contain. Where and how is this information to be stored? The assumptions that only irregular or unpredictable information goes in the lexicon and that the lexicon respects some modular boundaries, are at variance with evidence of richer and earlier interactions in word processing. For example, priming effects of false derivatives (e.g. the word corner priming its apparent base corn in English, baguette ‘stick’ priming bague ‘ring’ in French, or gallone ‘gallon’ priming gallo ‘rooster’ in Italian) are incompatible with the idea (Pinker & Ullman 2002) that a step of lexical look-up always precedes morphological segmentation rules in word processing, and that successful look-up (i.e. access of a fully stored word) inhibits segmentation. At the same time, evidence of fall priming fell (and the parallel failure of tell to prime tall) suggests that morphemic segmentation is not the only source of priming, and that word processing crucially involves co-activation of paradigmatically-related non-target forms. In fact, co-activation effects in word processing appear to be more global than pair-wise. They are modulated by the size and distribution of the lexical members of entire word families such as inflectional paradigms or derivational families.
Quantitative evidence of this interdependency and its impact on processing comes from work by Milin and colleagues (2009a, 2009b), investigating the facilitatory effects of intra-paradigmatic and inter-paradigmatic entropy on visual word recognition. These distributional effects accord well with evidence that time latencies in the human processing of words out of context are a function of so-called lexical uniqueness points, i.e. those points in the input sequence where the target input form can be distinguished from the set of all onset-overlapping words (Balling & Baayen 2008, 2012). From an acquisitional perspective, it would be very difficult to reconcile interactive effects of this kind with a modular view of the processing architecture, segregating sublexical constituents from combinatorial rules yielding fully inflected forms.

All this evidence supports the view that the paradigmatic (or vertical) organization of word forms in the mental lexicon is interdependent with the syntagmatic (or horizontal) linearization of sublexical constituents. Given an unsegmented input sequence of sounds or orthographic symbols, speakers learn to make context-sensitive predictions that approximate the conditional probabilities of an upcoming symbol given a series of preceding symbols. Conditional probabilities depend on the frequency distribution of all stored sequences with overlapping onsets. Since regularly inflected forms, like walk, walks, walked and walking, tend to share the same stem across all paradigmatically-related forms, their conditional probabilities tend to be less skewed on the stem-ending transition. This leads to an increase in prediction error at those boundaries, and provides evidence for implicit structure between successive sublexical constituents.

2.2 Evidence from computational modelling

The advent of connectionism in the 80’s popularised the idea that the lexical processor consists of a network of parallel processing units (functionally equivalent to neuron clusters) selectively firing in response to sensory stimuli (McClelland & Elman 1986; Norris 1994; Luce & Pisoni 1998). In processing the input stream, sensory information initiates concurrent activation (firing) of appropriate nodes, which respond to features/units of the input, as the input unfolds. Co-activation makes nodes compete with each other for selection, until an optimal candidate (generally, the most highly activated one) is singled out as the final winner.

Another important feature of connectionism is that it explicitly rejects the idea that word recognition and production require a dichotomous choice between storage (when the target word is already fully memorised) and processing (if it is to be assembled on-line from its constituent parts). In classical multi-layered perceptrons, the mapping of an input full form onto its morphological constituents is a continuous function of the statistical regularities in the form and meaning of different words. Since form-meaning mapping is predicted to be a graded phenomenon, perception of morphological boundaries in experimental data may vary as a result of the probabilistic support they receive from frequency distributions of acquired exemplars (e.g. Plaut & Gonnerman 2000; Rueckl & Raveh 1999; see also Hay & Baayen 2005).

Nonetheless, early connectionist models did not deal with many aspects of word processing in a satisfactory way. To begin with, the intrinsic temporal nature of the language input, unfolding and changing human expectations through time, was modelled in a rather ad hoc way. On the input layer, specific nodes were to be wired-in
to selectively respond to different instances of the same symbols in different temporal contexts. To illustrate, for a connectionist network to respond to \( n \) in knot, a specific \( k_No \) node (meaning a node fired by \( n \) when it is preceded by \( k \) and followed by \( o \)) has to be built in. As input nodes are not trained, the immediate implication of this prior encoding is that an input layer for Italian (where \( n \) cannot possibly be preceded by \( k \)) could not be used to deal with English.

Secondly, early connectionist simulations typically modelled word acquisition as the mapping of a base input form onto its inflected output form (e.g. \( go \rightarrow went \)). As observed by Baayen (2007), this protocol is compatible with the view of a redundancy-free lexicon, and adheres to a derivational approach to morphological competence, reminiscent of classical generative morphological theories. As we saw above, the view of a redundancy-free, derivation-based lexicon, however, is poorly motivated, unsupported by empirical evidence and logically unnecessary.

Thirdly, gradient descent protocols for training neural networks require supervision, i.e. the idea that the learner is systematically shown the correct output (e.g. went) for a certain input (go) during acquisition. Unless the network knows the expected response, it is in no position to adjust its connection weights to and from the hidden layer and learn the response. Now, we have scant evidence that children are supervised in this way: even when they are corrected (which is not a systematic event) they do not appear to take much notice of it.

Later connectionist architectures have tried to address all these open issues. In particular, recurrent neural networks (Rumelhart et al. 1986) have offered a principled solution to a) the problem of representing time, and b) the problem of learning without supervision. In simple recurrent networks (Elman 1990; Jordan 1986), the input to the network at time \( t \) is represented by the current level of activation of nodes in the input layer (as in classical connectionist networks) augmented with the level of activation of nodes in the hidden layer at the previous time tick \((t-1)\). In this way, the network keeps track of its past activation states and develops a serial memory of previous inputs. Many tasks have been used with recurrent networks, among which a simple but very powerful one is predicting the upcoming symbol at time \( t+1 \), given previous exposure to a sequence of symbols up to time \( t \). Simulating prediction is very ecological, since prediction is known to be heavily involved in human language processing (Altmann & Kamide 2007; DeLong, Urbach & Kutas 2005; Pickering & Garrod 2007). In addition, it provides a natural way to get instantaneous, observable feedback. All the network has to do is to wait for the upcoming symbol to show up. If it is different from what was expected and predicted by the network, the network parameters are adjusted for the currently presented input to be the network’s safest guess given the past evidence.

Along the same lines, Temporal Self-organizing maps (TSOMs, Ferro et al. 2011; Marzi et al. 2014; Pirrelli et al. 2015) have recently been proposed to model the dynamic topological organisation of memory nodes selectively firing when specific symbols are input to the map in specific temporal contexts. A temporal context is loosely defined as a temporal slot (position) in a time series of input symbols, or a window of surrounding symbols. Context-sensitive node specialisation is not wired in the map’s connections at the outset (as in traditional connectionist models), but it is something that emerges as a function of input exposure in the course of training (Marzi et al. 2016). High-frequency input sequences develop deeply entrenched connections and highly specialised nodes, functionally corresponding to human expectations for possible continuations. Low-frequency input sequences tend to fire blended node
chains, i.e. sequences of nodes that respond to a class of partially overlapping sequences. This is what distinguishes holistic, dedicated memorisation of full forms from chunk-based storage of low-frequency forms, sharing memory chunks with other overlapping forms.

TSOMs suggest a plausible, ecological way to conceptualise child word learning. Accordingly, children store all words they are exposed to, no matter whether they are regular or irregular, morphologically simple or complex. In addition to that, the self-organisation of items in the mental lexicon tends to reflect morphologically natural classes, be they inflectional paradigms, inflectional classes, derivational families or compound families, and this has a direct influence on morphological processing.

This conceptualisation has important theoretical implications on what constitutes the morphological competence of a human language learner. On the one hand, surface word relations represent a fundamental domain of morphological competence. On the other hand, in all languages, words happen to follow a Zipfian distribution in the learner’s input, with very few high-frequency words, and a vast majority of exceedingly rare words. Thus, word paradigms happen to be attested only partially also for high-frequency lexemes (Blevins et al. 2017), and learners must then be able to generalise available knowledge, and infer the inflectional class to which a partially attested paradigm belongs, for non-attested cells to be filled in appropriately. This is called the cell-filling problem (Blevins et al. 2017; Ackerman & Malouf 2013; Ackerman et al. 2009); given a small set of attested forms for a given paradigm, the learner has to guess all other missing forms.

In what follows I will first provide an informal introduction to principles of Hebbian learning (section 3), and then move on to consider the implications of these principles for a model of discriminative word learning based on co-activation/competition effects in lexical storage and processing (section 4). Finally, I will check the consequences of this view for a theory of the mental lexicon (section 5).

3. TSOMs and Hebbian Learning

The core of a TSOM consists of an array of nodes with two weighted layers of synaptic connectivity (Figure 1). Input connections link each node to the current input stimulus (e.g. a letter or a sound), which is represented as a vector of values in the [0, 1] interval, shown to the map at one time tick. Temporal connections link each map node to the pattern of node activation of the same map at the immediately preceding time tick. In Figure 1, these connections are depicted as re-entrant directed arcs, leaving from and to map nodes. Nodes are labelled by the input characters that fire them most strongly. ‘#’ and ‘$’ are special characters, marking the beginning and the end of an input word respectively.

Each time \( t \) an input stimulus is shown (e.g. an individual character or a phonological segment in a word), input activation propagates to all map nodes through both input and temporal connections (short-term processing) and the most highly activated node, or Best Matching Unit (BMU), is calculated. Following this short-term step, node connections are made increasingly more sensitive to the current input symbol, by getting their weights \( w_{i,j} \) (from the \( j \)-input value to the \( i \)-node) closer to the current input values \( x_j(t) \). The resulting long-term increment \( w_{i,j}(t) \) is an inverse function of the topological distance \( d_i(t) \) between each node and the current BMU\( (t) \), and a direct
function of the map’s learning rate $I(E)$ at epoch $E$. $I(E)$ is a dynamic parameter that decreases exponentially with learning epochs to define how readily the map can adjust itself to the input.

**Figure 1**: Functional architecture of a Temporal Self-Organising Map (TSOM). Shades of grey represent levels of activation in the map nodes after the word #pop$ is input. Directed arcs represent forward temporal connections between consecutively activated nodes.

In particular, given the input bigram ‘AX’,

(i) the connection between BMU(‘A’) at time $t-1$ and BMU(‘X’) at time $t$ is strengthened (entrenchment);

(ii) the connections to BMU(‘X’) from all the other nodes are weakened (competition).

The two principles, based on Hebbian modelling of mechanisms of synaptic plasticity by adaptation to input through learning (Hebb 1949), are also strongly reminiscent of Rescorla & Wagner’s (1972) discriminative equations.

For our present purposes, it is important to emphasise that the interaction between entrenchment and competition accounts for effects of context-sensitive specialisation of map nodes for input strings. If the bigram ‘AX’ is repeatedly input to a TSOM, the map tends to develop a specialised BMU(‘X’) for ‘X’ in ‘AX’ and a highly-weighted outward connection from BMU(‘A’) to BMU(‘X’). Since node specialisation propagates through time, a TSOM is thus biased in favour of memorising input strings through BMUs structured in a word-tree, as opposed to a word-graph (Figure 2).
Figure 2: When trained on the three mini-paradigms ‘AX’, ‘AY’, ‘BX’, ‘BY’, ‘CX’, ‘CY’, a TSOM tends to progressively move away from a graph-like allocation of nodes to symbols (left panel) towards a tree-like allocation (right panel). The extent to which context-sensitive specialisation takes place is a function of intra-paradigmatic and inter-paradigmatic word distributions (see main text for details).

As a general point, context-sensitive specialisation of BMUs allows a TSOM to allocate specific resources to input symbols that occur in specific temporal contexts. In this way, a TSOM develops a growing sensitivity to surface distributional properties of input data (e.g. language-specific constraints on admissible symbol arrangements, as well as probabilistic expectations of their occurrence), turning chains of randomly connected, general-purpose nodes into specialised sub-chains of BMUs that respond to specific letter strings in specific contexts. This ability is fundamental to storing symbolic time-series such as words.

4. Lexical co-activation/competition effects of memory self-organisation

Word co-activation and competition are known to account for effects of family size and frequency of neighbouring words in a variety of word processing tasks. Visual recognition of words that are surrounded by a large number of neighbours is known to be facilitated: printed words with many neighbours are recognised more quickly than words with fewer neighbours (see Andrews 1997 for a review). However, facilitation is modulated by the frequency distribution of neighbours, with a low-frequency target word suffering from the presence of a high-frequency neighbour. A reversal from facilitation to inhibition is shown in spoken word recognition (Luce & Pisoni 1998; Magnuson et al. 2007), where many neighbours are found to delay (rather than speed up) recognition of a target word.

Chen and Mirman (2012) attribute reversal to fundamental processing differences in the two tasks. In written word recognition, the units (nodes) fired by the letters making up the target word are fully activated simultaneously. Other possible non-target neighbours are co-activated too, with co-activation of shared units speeding up, and ultimately facilitating, visual word recognition. Competition at uniqueness points (where overlapping neighbours become perceptually distinct) is weak, since nodes of the target word are fully activated by all printed letters in input, whereas nodes of non-target competitors are only partially activated by the same letters. In the end, facilitatory co-activation outweighs the inhibitory effects of competition.
Conversely, in spoken word recognition, stimuli are presented in a strictly serial order. At any point \( t \) in processing time, all non-target candidates sharing the same onset up to \( t \) are activated equally strongly, and compete for primacy on a par with the target word. This causes more processing uncertainty at uniqueness points, where more alternative continuations of the shared onset are simultaneously entertained with comparable strength. In this scenario, the inhibitory effects of neighbour competition outweigh facilitation.

The difference is pictorially illustrated in Figure 3. In the figure, the two graphs represent levels of processing responses (activation) of nodes in a TSOM reading a written Greek word (Figure 3a), and in a TSOM recognizing the corresponding spoken word (Figure 3b). In both cases, the input word (whether written or spoken) is \( \alpha\gamma\alpha\pi\omega \) ‘I love’. A double-circled node in both graphs indicates the currently activated node. Note that, in Figure 3a, not all possible continuations of the string \( \alpha\gamma\alpha\pi\omega \) are equally activated, but the node corresponding to the letter \( \omega \) shows a higher activation level than other nodes do, due to the corresponding written stimulus being present in the input. Conversely, upon hearing [p] in [aGap] the network will project its expectations on different possible continuations of [aGap], whose strength is a nonlinear function of their distribution frequency (due to principles of Hebbian learning). If all continuations are equally likely, activation levels of their corresponding nodes will tend to be equally high. Thus, the increase in processing uncertainty goes up with the number of available alternatives and their levels of activation. Ultimately, the support of compatible competitors is outweighed by their inhibitory effect on word processing.

Figure 3: a) Levels of processing response by nodes in a TSOM upon reading the letter \( \pi \) in \( \alpha\gamma\alpha\pi\omega \) ‘I love’; b) levels of processing response by nodes in a TSOM upon perceiving the sound [p] in the same spoken word.
4.1 Word recoding and word recall

How can TSOMs model effects of lexical co-activation and competition in the mental lexicon? TSOMs can be used to simulate a number of word-processing tasks. In particular, the online response of a TSOM processing a time series of symbols presented to the map one at a time provides a good example of online serial behaviour. At each step, all map nodes fire concurrently, but only the most highly activated one, or *Best Matching Unit* (BMU) wins over the others. In processing an input word, levels of activation of nodes responding to each input symbol are integrated on the map over time ticks, thus forming a distributed representation of the map’s processing history, with BMUs of past symbols showing peaks of maximum activation. Such an integrated pattern of activation can be seen as the distributed result of serial word recoding. Notably, memory of past symbols, encoded as distributed activation peaks, shapes the maps expectation for incoming symbols.

Another ecological task consists in the immediate recall of an input word from the integrated pattern of node activation in a map just being exposed to that word. As all BMUs are simultaneously activated in the integrated pattern, word recall is a good example of processing a parallel input.

In serial word recoding, we can monitor the online expectations of a map for an upcoming node to be activated, while the map is processing an input word. At each processing step, the map updates its internal state of node activation and singles out the current BMU (i.e. the maximally activated node). Then, it updates its expectations for an upcoming stimulus by propagating the current state of node activation through its forward temporal connections before the new stimulus is shown. Such a prediction-based pre-activation of map nodes is a form of context-sensitive priming of the map, which may successfully anticipate the next input stimulus. We can measure how successful the map is in anticipating any input word at any point in time, by averaging the number of hits (i.e. correctly predicted symbols) by the length of the input sequence.

Figure 4a shows the predictive bias of a map trained on Greek verb forms (Bompolas et al. 2016, 2017), as an input word is processed from left to right. $x = 0$ on the horizontal axis marks the word position immediately following the base verb stem. The three lines illustrate how prediction increases for three different classes of allomorphy in Modern Greek aorist verb stems (Ralli 2005, 2006): asigmatic allomorphs, e.g. *pern-o* ‘I take’ ~ *pir-a* ‘I took’ (dotted line), phonological sigmatic allomorphs, *dulev-o* ‘I work’ ~ *đulep-s-a* ‘I worked’ (dashed line), morphological sigmatic allomorphs, e.g. *aγαρ(a)-o* ‘I love’ ~ *aγαπ-s-a* ‘I loved’ (solid line). Although sigmatic morphological and phonological allomorphs appear, on average, easier to predict, asigmatic allomorphs present a steeper increase in prediction as more of the input word is shown to the map. In other words, inflected forms containing asigmatic stem allomorphs (traditionally classified as irregular) become increasingly easier to process in their final part than regular forms do.
Figure 4: Top: marginal plot of interaction effects between word position (x axis) and classes of inflectional regularity (solid line = sigmatic morphological stems, dashed line = sigmatic phonological stems, dotted line = asigmatic stems) in an LME model fitting average BMU prediction (y axis) in a TSOM trained on Modern Greek verbs. Bottom: the same interaction effects are plotted separately for the stem (in the [-6, -1] range of x values) and the inflectional ending (in the [0, +5] range of x values). Lines are shown in the range of attested data (adapted from Bompolas et al. 2017).

Figure 4b shows why it is so. For each class of Greek stem allomorphy, the figure provides two distinct plots, one for stems and one for endings. Irregular forms (those containing asigmatic stem allomorphs) are systematically more difficult to predict at the stem level (lines in the [-8, -1] range of x values). However, they become easier to predict over their inflectional endings (lines in the [0, +5] range of x values). The reversed pattern obtains for regular forms (i.e. those containing phonological and morphological sigmatic stems), whose easy-to-predict stems are followed by less predictable endings.

A drop in prediction rate between the end of the stem and the start of the suffix is what is expected for highly inflecting morphologies, due to the variety of endings that can follow the morpheme boundary. We suggest that this level of uncertainty is diagnostic for structural complexity: a drop in the level of expectation for an upcoming node at the morpheme boundary is the effect of a perceived morphological discontinuity on the temporal connectivity of the map, in line with distributional accounts of morphological structure (Hay 2001; Bybee 2002; Hay & Baayen 2005). Due to stem
allomorphy, irregulars are more difficult to predict at the beginning, where more formal variation is expected, as illustrated by the word graph of Figure 5. Once the stem allomorph is selected, however, the range of admissible endings is much more constrained than the range of endings in regulars. This causes less uncertainty and a reduced level of perceived morphological structure.

In word recall, the map’s initial state of activation is the integrated activation pattern (IAP) resulting from exposure to an input word $w$. Each node in the IAP is associated in traversing the IAP to reinstate, in the appropriate order, the sequence of symbols making up $w$. This is done by first activating the start-of-word node, to then propagate its expectations through forward connections. Expectations are summed up with levels of node activation in the IAP, and the resulting maximally activated node (or BMU) is calculated. The output of the map is the symbol associated with its current BMU. The process is repeated over again for each symbol, until the end-of-word node is reached. The input word $w$ is recalled, if all its symbols are reinstated in the appropriate order. Otherwise, recall fails.

![Figure 5: Word graph representation of a few forms of perno. Stem allomorphs are highlighted in grey. Double-circled nodes represent word final states.](image)

Figure 6 shows the earliest learning epoch since regular forms (solid line) and irregular forms (dotted line) are recalled correctly, plotted as a function of their length. As a general trend, longer words are recalled later than shorter words. This is a “buffering” effect, showing that longer words are recalled less easily than short words are. In fact, for longer words, more symbol representations have to be maintained concurrently while being manipulated for recall, and this raises the chances of substitution, deletion and transposition errors. On the other hand, for longer forms, regulars are recalled earlier than irregulars are, as shown by learning epoch values in the $[9, 14]$ length range, where values of regulars are systematically lower than those of irregulars. We take this as evidence that regulars are recalled more easily than irregulars, in spite of the latter being more easily predicted than the former. This is a morphological regularity effect. In regular paradigms, all inflected forms share the same stem (Figures 3a and 3b), and recall can take advantage of their joint support. When stems are long enough, their positive support makes it up for the interference caused by their co-activation. This is not the case for regular forms with shorter stems, which are learned comparatively later than irregular forms with shorter stems are.
Moving to a different domain of morphological inquiry, analysis of compounds offer another interesting case of human perception of structural discontinuity. Ferro and colleagues (2016) discuss a key question concerning the representation of compound words in the mental lexicon: whether and how morphological structure plays any role in the way they are stored and accessed. In particular, they focus on evidence from written production: namely, elevation in typing time for the initial letter of the second constituent of an English Noun Noun compound (e.g. b in snowball), relative to the typing time for the final letter of the first constituent (w in snowball) (Gagné & Spalding 2016). This elevation indicates that production of compounds differs from that of monomorphemic words, and that the semantic transparency of the two constituents (i.e. the extent to which the semantics of the whole compound is a function of the semantics of its constituents) plays a role in this. At the same time, similar elongation effects are found in so-called pseudo-compounds like carpet (apparently made up out of two words: car and pet), which are reminiscent of effects of automatic morphemic segmentation of apparent derivatives (e.g. corner or baguette). Once more, neither pre-lexical (Taft & Forster 1975) nor post-lexical (Giraudo & Grainger 2000) accounts of morphological structure can account for the entire range of effects reported here. The evidence appears to elude also so-called race models (Schreuder & Baayen 1995), which posit two parallel pathways for compound processing, one for holistic and the other for compositional interpretation. In a mono-morphemic word like carpet, for example, the whole word path should take precedence, but this would cancel out effects of apparent compositionality.

TSOMs can account for distributed processing effects at the peripheral level of lexical access (Ferro et al. 2016). Figure 7a plots ease of prediction across compounds (solid line) and pseudo-compounds (dashed line) by letter distance to the morpheme boundary between the two (real or apparent, respectively) compound constituents C1 and C2 (with x = 0 corresponding to the position of the first letter of C2). The steeper prediction rate for pseudo-compounds is in line with evidence of faster speedup rate in typing of mono-morphemic vs. compound words by human subjects. Like irregularly inflected items, pseudo-compounds suffer from competition by onset sharing...
neighbours, as their apparent first constituents (e.g. car in carpet) are usually comparatively short and confusable. This is what we would expect: pseudo-compounds are, in fact, only apparent compounds, i.e. mono-morphemic words, whose syllables just happen to be formally identical to independent words. Nonetheless, as soon as the major source of processing uncertainty is successfully dealt with, i.e. when the apparent constituent boundary is reached, the prediction rate grows more quickly and the final part of the mono-morphemic input word is more easily processed. This is confirmed by the plot of Figure 7b, showing a drop in prediction rate between C1 and C2 for both real compounds and pseudo-compounds, in line with Gagné and Spalding’s evidence. In addition, the plot shows a higher level of uncertainty in processing the C2 of compounds, as opposed to pseudos, and a steeper increase in processing the C2 of pseudo-compounds than the C2 of real compounds. The trend is in keeping with evidence of a significantly stronger perception of structural complexity in compounds than pseudo-compounds, by human subjects.

![Figure 7a](image1.png)

**Figure 7a:** Marginal plots of interaction effects between compounds vs. pseudo-compounds and letter distance to morpheme boundary in an LME model fitting anticipation of up-coming BMUs by a TSOM. Negative and positive x values indicate letter positions located, respectively, in the first and second constituent. Anticipation is plotted across whole (pseudo)compounds (top panel), and by individual constituents (bottom panel) (adapted from Ferro et al. 2016).
5. General discussion

TSOMs represent, primarily, general models of serial memories. However, there is more to TSOMs than just storage. First, they can simulate effects of structure-sensitive global self-organisation of concurrently stored word forms. In TSOMs, competition and co-activation of neighbouring words are grounded in shared memory/processing resources. This means that the map representations of two word neighbours tend to share a pool of identical memory nodes (or topologically close nodes, see Figure 3 above), and explains why activation of one lexical representation entails concurrent activation of another lexical representation.

Due to Hebbian specialisation, this dominant effect is modulated by word frequency. A high-frequency word tends to recruit dedicated nodes, i.e. memory nodes that are selectively fired by that word. This does not cancel out co-activation of other neighbouring lexical representations, but considerably reduces the effect magnitude.

Competition is another name for co-activation. What has a facilitatory effect on some processing tasks, e.g. concomitant support of co-activated neighbours in parallel processing of simultaneously presented stimuli, may have an inhibitory impact on other tasks. For example, in prediction-based processing of serially presented stimuli, the concomitant activation of a high-frequency, deeply entrenched non-target representation may considerably interfere with recognition of the target representation, thus turning neighbouring friends into potential “enemies”.

A paramount linguistic effect of Hebbian specialisation is reduced perception of morphological structure. In a TSOM, morphological structure is interpreted as a less skewed expectation for an upcoming stimulus at morpheme boundaries, leading to an increase in prediction errors at those boundaries. This makes less frequent forms more difficult to predict. At the same time, however, they exhibit an implicitly segmented structure, which is conducive to induction of novel forms. Conversely, more frequent words are easier to predict and process, but their internal structure is less salient perceptually. We illustrated these effects in connection with high-frequency irregularly inflected verb forms and so-called pseudo-compounds.

For our present concerns, the most important feature of TSOMs is that they implement, in a straightforward, algorithmic way, a non-trivial interdependency between memory and processing (Marzi & Pirrelli 2015). Their lexical representations (or integrated memory traces) appear, in a sense, to cache the map’s processing expectations. The nodes making up the lexical memory trace of a word form are exactly those units that are fired while the map is processing the word form. In addition, the processing behaviour of TSOMs depends on the long-term information stored in their nodes. The dependency is in fact two-way. On the one hand, processing is based on the temporary (short-term) re-activation of a pattern of memory nodes keeping (long-term) information in their connections. On the other hand, long-term storage is crucially shaped by the map’s processing behaviour, because it consists in recording successful, routinized processing responses of the map, prompted by recurrent input patterns. Looking at processing and memory from this perspective forcefully imposes the view that the two processes are, respectively, the short-term and long-term dynamic of a unique underlying process.
I believe that, in spite of its simplicity, this view is conducive to an insightful conception of the mental lexicon, whose main features (in a deliberately sharp contrast with the more traditional list of section 1 above) can be summed up as follows:

- **non-enumerative**: there is no finite list of stored items in the lexicon; there are many more (virtual) pathways between nodes than those entrenched through training; the notion of “wordlikeness” (or “lexicality”) is a gradient one (a lexical entry can be perceived by the map as more or less “typical”), and is not co-extensive with the notion of “being listed”
- **parallel**: items are activated simultaneously and accessed globally
- **dynamic**: information is never stable; every time a lexical representation is successfully accessed, its content changes accordingly (e.g. through consolidation of connection strengths); moreover, access of any lexical representation affects, more or less deeply, the activation state of all other representations in the same lexicon
- **processing-dependent**: the form of a lexical representation is fundamentally grounded in processing principles, and consists in the same processing units that are fired by the input word associated with the lexical representation
- **redundant**: it contains both lexical and sublexical structures
- **emergent**: word structure is not a prior, but it is the perceived by-product of stored, unsegmented input stimuli (full forms or units larger than full forms)
- **multidimensional**: the lexicon contains structural units defined on many different hierarchical levels, ranging from syllables and morphemes, to words and phrases, depending on the level of complexity of the input
- **two-way**: its content affects, and is crucially affected by processing

Taken all together, these points constitute the core defining features of an “integrative” view of the mental lexicon (Marzi et al. 2016), opposed to the more “discriminative” (or radically modular) views of language architecture that were dominant in the pre-connectionism years. It should be appreciated that TSOs represent only a possible implementation of this view. Other connectionist architectures can be envisaged that comply with the simple pool of Hebbian principles of section 3. It is this pool of principles (and their neurobiological correlates), we believe, that contributes and will contribute much to reshaping our way of thinking of language, its architecture and, ultimately, its ontological foundations. Pausing for thinking on the consequences of this change in perspective is also good for building novel theoretical frameworks or rediscovering old ones (Blevins 2016). Finally, it is bound to shed new light on time-honoured controversies in the linguistic literature, such as the morpheme-based vs. word-based nature of the morphological competence, or Hockett’s (1954) dichotomy between Item-and-Arrangement and Item-and-Process morphologies, by pointing to a principled, dynamic way to reconcile opposing views.

**Acknowledgements**

I wish to thank Angela Ralli for inviting me to the 4th Patras International Conference of Graduate Students in Linguistics. I enjoyed the event and benefited considerably from feedback from the
audience. Thanks are also due to Stavros Bompolas for his insightful feedback and support. Much of the experimental and computational work reported here is the fruit of a long-time productive collaboration with Claudia Marzi and Marcello Ferro. I am grateful to them for their invaluable contribution to the design, development and analysis of all computational models and data presented here.

References


