UDK 81'23
165.194
159.955:81
Izvorni znanstveni članak
Prihvaćeno za tisak 23. rujna 2014.

Claudia Marzi
Institute for Computational Linguistics – National Research Council, Pisa, Italy
claudia.marzi@ilc.cnr.it

Marcello Ferro
Institute for Computational Linguistics – National Research Council, Pisa, Italy
marcello.ferro@ilc.cnr.it

Emmanuel Keuleers
Ghent University, Belgium
emmanuel.keuleers@ugent.be

Perception of typicality in the lexicon: wordlikeness, lexical density and morphonotactic constraints

The extent to which a symbolic time–series (a sequence of sounds or letters) is a typical word of a language, referred to as WORDLIKENESS, has been shown to have effects in speech perception and production, reading proficiency, lexical development and lexical access, short–term and long–term verbal memory. Two quantitative models have been suggested to account for these effects: serial phonotactic probabilities (the likelihood for a given symbolic sequence to appear in the lexicon) and lexical density (the extent to which other words can be obtained from a target word by changing, deleting or inserting one or more symbols in the target). The two measures are highly correlated and thus easy to be confounded in measuring their effects in lexical tasks. In this paper, we propose a computational model of lexical organisation, based on Self–Organising Maps with Hebbian connections defined over a temporal layer (TSOMs), providing a principled algorithmic account of effects of lexical acquisition, processing and access, to further investigate these issues. In particular, we show that (morpho–)phonotactic probabilities and lexical density, though correlated in lexical organisation, can be taken to focus on different aspects of speakers' word processing behaviour and thus provide independent cognitive contributions to our understanding of the principles of perception of typicality that govern lexical organisation.

1. Introduction

Many aspects of lexical processing and organisation are shown to be related to issues of formal redundancy, i.e. to the extent to which surface word representations (either phonological or orthographical) overlap in the lexicon, which in turn appears to reflect some fundamental mechanisms of human memory for serial order. For example, the mental lexicon is known to store entries “in waves”, with confusable entries usually having similar beginnings and similar endings (e.g. *anecdote* vs. *antidote*), with initial sounds being more confusable in short words and final sounds being more confusable in longer words. The phenomenon appears to reflect primacy and recency memory gradients familiar from the literature on short–term and long–term memory (Aitchison 1987; Gupta 2005, 2009, among others).

In fact, a large body of wordlikeness effects, reflecting the extent to which a particular string of symbols is perceived as typical of a given lexicon (Bailey & Hahn 2001), appear to interact with memory issues, and, in particular, with the encoding of time–series of symbols in the long–term lexical storage. In the present paper, we intend to show that computational models of serial memories can provide a principled account of at least some of these effects, by grounding them on some basic mechanisms of co–activation and competition between concurrently stored words.

Most psycholinguistic models of the mental lexicon are based on the fundamental hypothesis – confirmed by neuro–functional evidence – 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 of the appropriate nodes that respond to features/units of the input as they unfold through time. When the activation spreads to the lexical level, multiple lexical candidates are co–activated and compete with each other for final selection. Goodness–of–fit criteria guide the activation towards the optimal candidate, which is eventually singled out as the final winner. These basic assumptions appear to capture aspects of the dynamicity of mental processes (e.g. Bybee 1995) and mimic a plausible architecture of the neurobiological substrate (e.g. Hickok & Poeppel 2004).

Different activation–based models make different claims as to the number and nature of the non–target words that are partially co–activated as a result of the network being exposed to a particular target input word. For example, the *cohort model* in its original form (Marslen–Wilson 1987) makes the prediction that non–target words are activated as a direct function of left–to–right auditory overlap in speech perception. Due to the serial nature of both input processing and input stimulus, non–target words are (partially) activated if they share a conspicuous auditory onset with the target input stimulus, thus putting considerable emphasis on the perceptual salience of word–initial sounds in lexical speech recognition.

A different approach to competitor neighbourhood is proposed by Luce and Pisoni (1998), based on the notion of minimal editing distance from the target input: a non–target word is considered to be part of the neighbourhood of a target word, if the former can be obtained from the latter by a process of sound/letter substitution, deletion or insertion. Both neighbourhood size and frequency distribution of neighbours are taken to play a role in the resulting competition. Target words selecting many similar–sounding neighbours are recognised less accurately and less quickly than words occurring in sparse or less confusable neighbourhoods. These competition effects are also mediated by individual frequency of neighbours: high–frequency neighbours are likely to compete more strongly with the target word thus reducing recognition performance, whereas low–frequency words tend to exert a less perceivable influence.

Both cohort–based and neighbourhood–based effects (and variants thereof: the TRACE model, McClelland & Elman 1986; the Shortlist model, Norris 1994) are shown to be observable in auditory speech recognition (Mirman et al. 2010), but with a different timing. Since cohort effects are more related to on–line left–to–right word processing, they tend to emerge at early processing stages, thereby accounting for the human capacity to PREDICT candidate input words on the basis of the already processed input stream. On the other hand, since neighbourhood membership is more based on holistic similarity and relatively unconstrained by the order and position of incoming signals, neighbourhood density has a stronger influence on WORD SELECTION rather than WORD PREDICTION. It remains to be seen whether these apparently contradictory effects are the result of independent processes, or rather the time–bound, dynamic manifestation of a single underlying lexical co–activation state.

It is important to emphasise that performance–oriented evidence (such as frequency–based effects or effects of neighbourhood size on speech recognition) can shed light on the memory–based representations forming part of the speaker's lexical knowledge. In addition, theoretical assumptions about long–term lexical representations, such as the conjunctive view that a word's form consists of segmental units representing sound identity and position in a unitary manner, predict that non–target words that share segments in the same/similar position are co–activated in lexical processing. This means that representations and processing are mutually implicated. In this perspective, Temporal Self–Organising Maps (TSOMs) provide a computational framework to test fundamental mechanism underpinning serial cognition and for simulating processes of lexical organisation. They represent grids of processing nodes that mimic the spatial and temporal organisation of memory structures supporting the processing of symbolic sequences. We believe that computational modelling can help us gain considerable insights into the mutual relationship between representation (memory) and processing strategies (perception), and empirically verify, under controlled simulations of word stimuli, that memory structures represent the way external stimuli are processed and perceived.

2. TSOMs

The TSOM architecture consists of a grid of topologically organised memory nodes, representing one layer of neurons, with dedicated sensitivity to time–bound stimuli.

In a variant of the classical Kohonen architecture, SOMs augmented with re–entrant Hebbian connections defined over a temporal layer can encode probabilistic expectations upon the incoming stimulus (Koutnik 2007; Ferro et al. 2010; Pirrelli et al. 2011; Ferro et al. 2011; Marzi et al. 2012a, 2012b). Temporal first–order connections, providing the state of activation of the map at the immediately preceding time step, can be interpreted as encoding the map's probabilistic expectations of up–coming events on the basis of past experience, making room for memorising time series of symbols as activations chains of nodes.

Upon presentation of an input stimulus, all map nodes are activated synchronously, but only the most highly activated one, the so–called *Best Matching Unit* (BMU), wins over the others. Neighbouring nodes become increasingly sensitive to input pattern (e.g. symbols, letters, etc.) that are similar in both encoding and distribution through training.

Each node in the map is connected with all elements of the input layer through communication channels with no time delay, whose strength is modified through training (see Figure 1). TRAINING consists in showing the map one input form at a time, each sampled according to its distribution.

Words can be represented as temporal patterns (strings of acoustic/written symbols), and are produced by and input to a TSOM one symbol at a time on the input layer. The input layer is implemented as a binary vector, with one pattern of bits uniquely associated with each symbol. A whole word is presented to a map starting with a start–of–word symbol ('*#*') and ending with '*$*'. Map's temporal activation state is initialised upon presentation of a new word. This implicates that the map's activation state upon seeing the current input word form has no short–term memory of past word forms. Nonetheless, the map's overall state is affected by previously shown words through long–term learning (STORAGE).

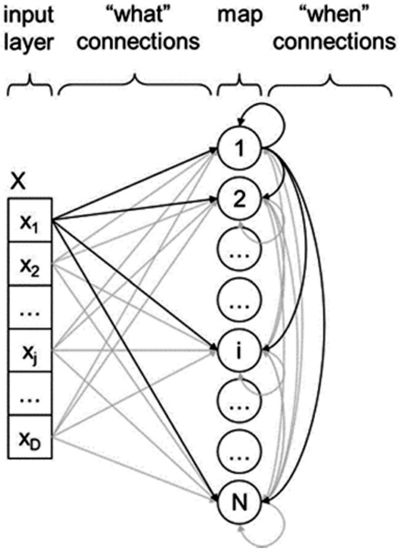


FIGURE 1: Outline architecture of a TSOM. Map nodes present two levels of connectivity on both spatial (*what* connections) and temporal (*when* connections) representation levels.

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 and its temporal context (e.g. a 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. Due to this training dynamic, selective specialisation of nodes is the natural bias of a TSOM.

On the other hand, activation spreading of WHEN connections through learning implements the map's propensity to accept novel forms by extending learned connections to local topological neighbourhoods. As sketched in Figure 1, WHAT connections define the communication channel between the input layer and the map proper, whereas each map node communicates with any other node through pre– and post–synaptic weighted connections, referred to as when connections.

In TSOMs, alternative chains of memory nodes may be co–activated by the same input sequence, but only one chain is preferred over the others during processing, depending on the degree of probabilistic support received by the network of long–term associative relations among stored word forms. In addition, since a TSOM stores recurrent processing steps through memory nodes and it uses them over again whenever possible, different chains of memory nodes may be co–activated by the similarly perceived input sequences. Accordingly, CO–ACTIVATION of memory nodes by different input words reflects the extent to which a TSOM perceives surface relations between words.

3. Quantitative correlates of wordlikeness in TSOMs

Bailey and Hahn (Bailey & Hahn 2001; Hahn & Bailey 2005) identified and investigated two quantitative correlates of the pre–theoretical notion of WORDLIKENESS, i.e. the extent to which a speaker perceives a de–contextualised sequence of symbols (either sounds or letters) as typical of her own native lexicon: PHONOTACTIC LIKELIHOOD and NEIGHBOURHOOD DENSITY. Phonotactic likelihood defines the estimated probability that a particular string results from the concatenation of smaller symbol chunks (n–grams), whose distribution is derived from the entire lexicon. Intuitively, the trigram 'AHR' is fairly likely to be found in a German lexicon (e.g. FAHREN), but would hardly be found at all as part of an Italian word. Hence, a string containing 'AHR' is bound to be judged/perceived as German–like, but it would be rejected as an Italian possible word, simply because of the presence of a trigram ('AHR') which has 0 probability to occur in typical Italian words.

Lexical neighbours are existing words that are similar to a specific target word. Procedurally, any such set can be obtained by altering the target word through insertion, deletion or substitution of one or more of the word's symbols. The more editing operations of this kind are allowed, the larger the resulting set of neighbours will be. For any given target word, one usually refers to the size of its set of neighbours as its neighbourhood density.

Phonotactic likelihood and neighbourhood density are strongly correlated notions. A target string containing high–frequency n–grams is likely to be surrounded by a large set of lexical neighbours. Conversely, some scholars have argued that frequency effects are in fact density effects in disguise: a large set of neighbours is likely to contain high–frequency words, and ultimately high–frequency n–grams (Bard 1990). Yet, Bailey and Hahn show that in spite of their being highly correlated, the two notions appear to independently explain wordlikeness judgement/perception, with neighbourhood density accounting for a larger part of their variance, and phonotactic probability explaining residual effects. Here, we would like to address the issue from a different perspective. Wordlikeness has not only to do with meta–linguistic awareness or acceptability judgements on strings, but it is assumed to play an important role in word processing and storage. As extensively discussed in the previous section, many accounts of language performance assume that in the course of word perception and production, the non–target lexical neighbours of an input word get co–activated by the current input stimulus, and become available for further processing stages.

Here we want to empirically assess the role of wordlikeness in the overall organisation of lexical items by TSOMs, with particular emphasis on issues of processing, storage, access, and recall.

We ran a series of simulations with four data sets: (i) a training dataset, composed by 700 German[1] fully inflected verb forms, obtained by selecting from CELEX (Baayen et al. 1995) the top 50 high–frequency paradigms and 15 inflected forms for each paradigm (the full set of present indicative (6) and präteritum (6) forms, the past participle, the infinitive and the present participle); (ii) for each of the 50 paradigms one verb form was not administered to a TSOM during training, and used to form a set of paradigmatically–related test words; (iii) an additional German test set was selected from CELEX, by picking up 25 more high–frequency paradigms forming a set of novel words. An additional test (iv) was designed to represent non–words to the TSOM trained on German verb forms. The test set was formed by selecting 50 Italian verb paradigms whose word forms should be perceived as unfamiliar or ill–formed from the TSOM perspective.

In Figure 2[2] we report mean neighbourhood values for each of the word sets i–iv. For each target word, its neighbourhood is defined as the set of the word forms (of data set i) lying at a 3–character editing distance from the target word. Neighbours at shorter distances than 3 were found to be too sparse to reach significance. For the same sets, we calculated word–averaged orthotactic probabilities (Figure 3). Both scores define a gradient of wordlikeness, ranging from most familiar (training) words (i) to least familiar ones (iv). Scores

[1] The choice of German as training language is made on grounds of availability and inflectional complexity. In particular, German verb morphology includes suffixation, stem alternation, circumfixation and combinations thereof.

[2] On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme datapoints the algorithm considers to be not outliers, and the outliers are plotted individually.

on test words (same verbs as in the training set but in different inflections) are shown not to be significantly different from training words. Our data also exhibit the expected correlation between size of the neighbour family of a target word, and its orthotactic probability (r=.41, p<.000005).

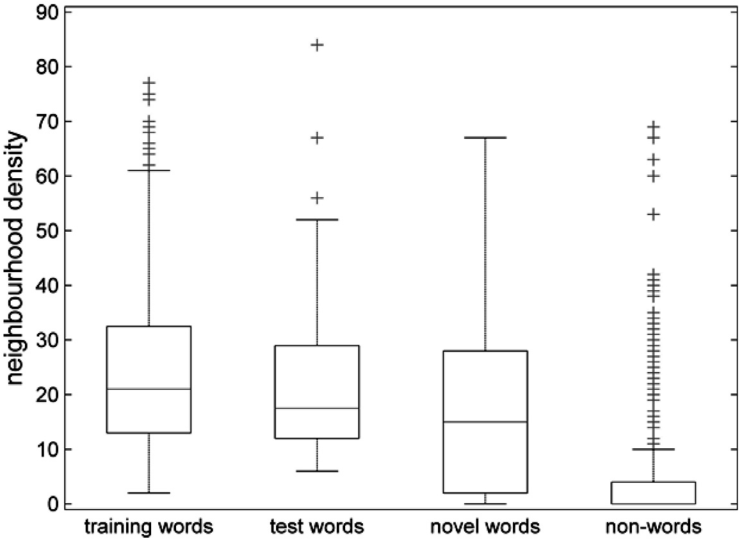


FIGURE 2: Neighbourhood density of words belonging to different datasets. The neighbourhood density of a word form is calculated as the number of training words within an edit distance of 3 from the given word form.

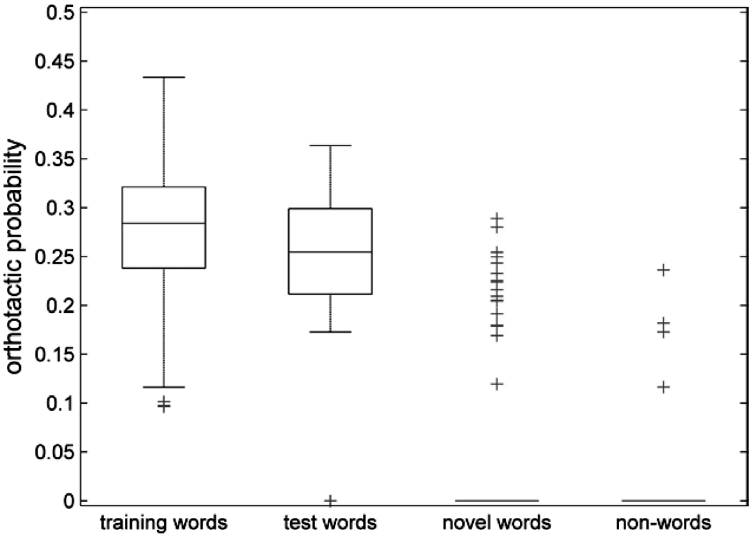


FIGURE 3: Orthotactic probability of words belonging to different datasets. The orthotactic probability of a word form is calculated as the geometric mean of conditional trigram probabilities (from training words) across the whole word form.

During training, input forms (i) were administered according to a function of their frequency distribution in CELEX, for 100 learning epochs. For all the four datasets we first evaluated the TSOM behaviour on two tasks: WORD RECODING and WORD RECALL. The former 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 recoded 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 '#' 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.

In Table 1, we give for each set of words recoding and recall accuracy percentage values (averaged over 10 different instances of the trained map). Standard deviation values refer to recall scores.

Data set	*Recoding accuracy*[3]	*Recall accuracy*	*Standard deviation*
training set (i)	100%	99.2%	0.6%
test words (ii)	100%	98.6%	1.9%
novel words (iii)	100%	51.0%	3.9%
non–words (iv)	100%	18.6%	1.8%

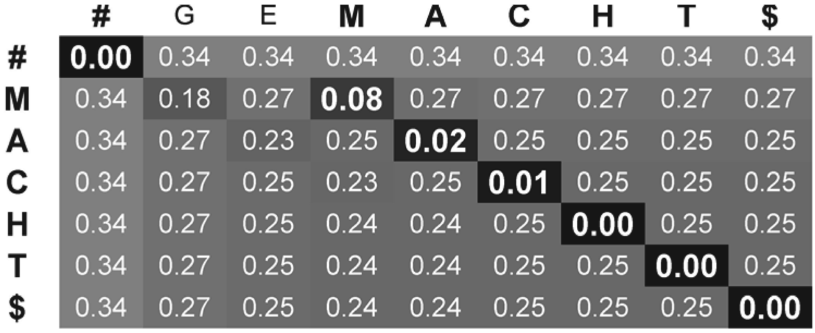
Table 1.

Note that accuracy values are very high in both recalling correctly word forms of the training set and its related test words. Accuracy decreases to 51% for German novel words (novel verb paradigms) and drops to 18.6% for Italian words (non–words from the TSOM perspective). Recall values strongly correlate with degrees of wordlikeness as estimated by both neighbourhood density (r=.582, p<.000005) and orthotactic probability (r=.733, p<.000005) scores.

[3] It should be appreciated that the representation of symbols adopted in the present work is orthographic and localist, meaning that it assumes independence (orthogonality) among all vector symbol codes on the input layer. An implication of this is that, from the map's perspective, each symbol is uniformly different from any other symbol. Uniform encoding means – for example – that letter 'a' in German and letter 'a' in Italian are assigned an identical binary vector on the input layer. Nonetheless, recoding the binary vector as a node activation pattern requires a process of context–sensitive perception of a symbol, which is internally represented by the map as a symbol with its context (conjunctive recoding). It follows that a map trained on the German lexicon would have little problems in recognising the Italian alphabet (the only potential difficulties arising in connection with language–specific diacritics) but it would perceive it with German orthotactic expectations. That's why 100% accuracy in Italian written word recognition by a map trained on a German lexicon is relatively unsurprising. Conversely, due to the map's system of orthotactic probabilities, word recall, which heavily depends on trained time–bound expectations, is deeply affected by language–specific conjunctive recoding.

It is important to emphasise that the notions of SERIAL ACTIVATION and PARALLEL CO–ACTIVATION of map nodes define two different dimensions of word-likeness. During training, node chains get specialised to respond to recurrent sequences of input symbols; this is obtained by strengthening weights over frequently–traversed inter–node connections. In recalling an input word, the map uses connection weights as predictions over upcoming symbols. Words are accurately recalled when the map can reinstate them as sequences of strongly expected chunks. Thus, the activation strength is a measure of the degree of the map's expectation for familiar sequences.

When more node chains are simultaneously activated by an input word, levels of co–activation define the extent to which the word is perceived as similar to other words in the lexicon. For example, when the German form GEMACHT is administered to the map, the BMUs responding to prefixed –MACHT are likely to be different from those responding to MACHT as a full form (Figure 4). Nonetheless, the latter nodes will be highly co–activated, eventually competing for selection with the more specialised GEMACHT nodes. This is relevant to the present discussion on two respects. First, differences in co–activation levels among competing nodes define perceptual distances between related words. An input word whose activation state strongly reverberates (i.e. is co–activated) with BMUs for other words is perceived as highly familiar. Secondly, levels of co–activations approximate levels of confusability between competing stimuli, thus weakening the map's expectation on upcoming symbols.



	#	G	E	M	A	C	H	T	$
#	**0.00**	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34
M	0.34	0.18	0.27	**0.08**	0.27	0.27	0.27	0.27	0.27
A	0.34	0.27	0.23	0.25	**0.02**	0.25	0.25	0.25	0.25
C	0.34	0.27	0.25	0.23	0.25	**0.01**	0.25	0.25	0.25
H	0.34	0.27	0.25	0.24	0.24	0.25	**0.00**	0.25	0.25
T	0.34	0.27	0.25	0.24	0.24	0.25	0.25	**0.00**	0.25
$	0.34	0.27	0.25	0.24	0.24	0.25	0.25	0.25	**0.00**

FIGURE 4: Co–activation distances for the input forms MACHT and GEMACHT.

We thus suggest that serial activation and parallel co–activation, however correlated, play distinct and potentially conflicting roles in the map's dynamic behaviour, the former favouring local recall, and the latter being more conducive to global lexical organisation effects and, as we will see in more detail below, to perception of structure.

To investigate this non–trivial interplay and its effects on the map states, we first measured activation levels of the map after exposure to each word from sets i–iv, by looking at the node–averaged activation strength of the integrated BMU chain responding to each input word[4]. The full form activation strengths for sets i–iv are shown as a box plot in Figure 5.

The difference in activation strengths is significant overall (Figure 5), with Figure 6 detailing the relative contribution of verb stems and verb endings to differences in activation strengths between training words (i) and test words in sets ii–iv.

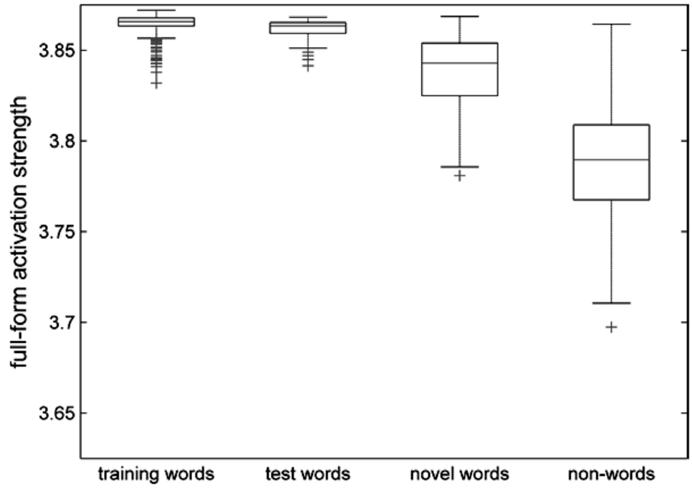


FIGURE 5: Activation strength of words belonging to different datasets. For each word form, the activation strength is calculated as the mean activation level of BMUs in the activation chain of the word form.

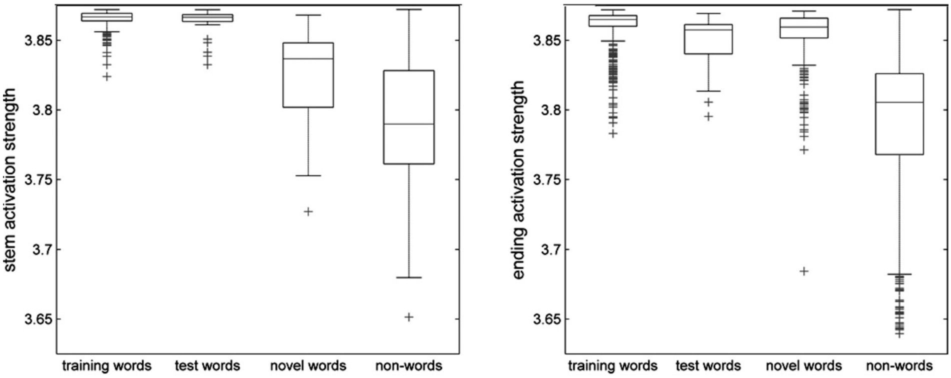


FIGURE 6: The activation strength of words belonging to different datasets calculated on word stems (left) and word endings (right) only.

[4] In training, node activation levels are normalised after each stimulus exposure for them to fit into a 0–1 activation range. To avoid ceiling effects, pure activation values (activation strengths) were averaged before normalisation. Note that we first averaged the activation level of each word across the 10 map instances, to then average word strengths across all words in each set.

A different measure of wordlikeness is offered by the node–averaged summation of inter–BMU connection weights for all words in the four sets (or connection strength, Figure 7). Connection weights measure the time–bound map's expectation for a certain node to be activated at the ensuing time step. A high connection strength means that the corresponding word was shown to the map a high number of times during training, or, alternatively, that it consists of high–frequency n–grams. Hence, it reflects orthotactic typicality to the extent the most typical n–grams are also the most frequent ones. The relative contribution of each word's stem and word's ending to whole–word effects in Figure 7 is provided in Figure 8.

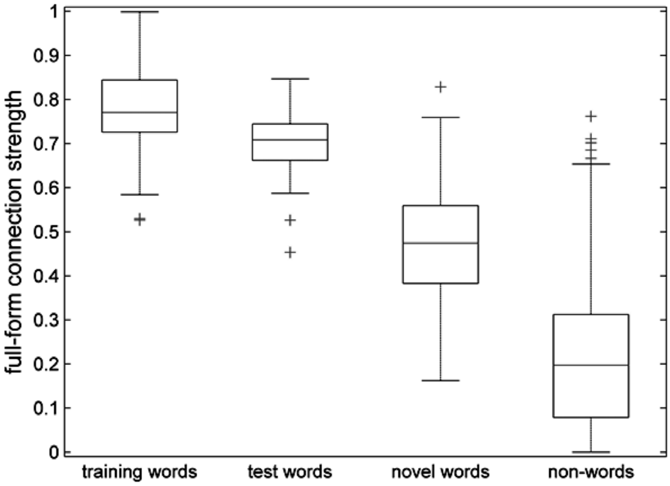


FIGURE 7: Connection strength of words belonging to different datasets. The connection strength of a word form is calculated as the mean strength value of the temporal connections between consecutives BMUs in the activation chain of the given word form.

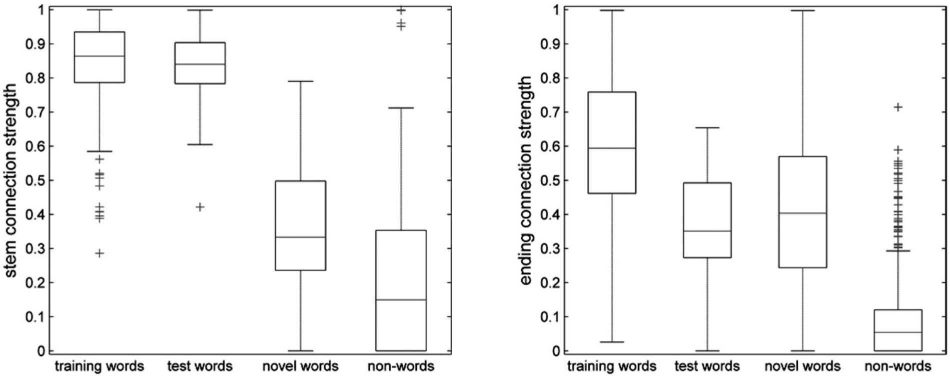


FIGURE 8: The connection strength of words belonging to different datasets calculated on word stems (left) and word endings (right) only.

Finally, we assessed how well the pattern of activation of each individual word in the four sets fit the overall integrated activation pattern for all words in the training set, calculated by cumulating over levels of activation of their BMU chains. We expected this goodness–of–fit measure to best reflect typicality of the activation chain left by a word on the map relative to the entire training lexicon. Results across all four datasets are shown in Figure 9.

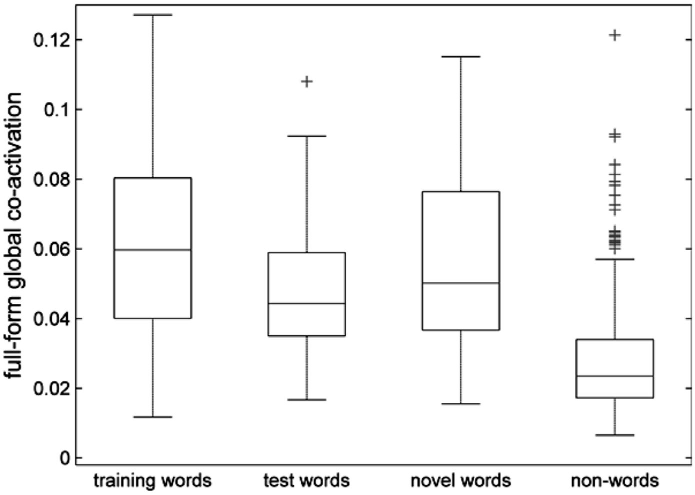


FIGURE 9: Global co–activation of words belonging to different datasets. For each word form, it is calculated as follows: i) an integrated activation pattern (IAP) is first obtained for each training word as the union of the activation patterns of the word's symbols; ii) IAPs of all training words are then cumulated into a global activation pattern (GAP); iii) for each word, its global co–activation is the node–averaged activation level of the word's BMUs in GAP.

4. A multi–factorial account of wordlikeness

Recently, Goldrick and colleagues (2010) offered a multi–factorial account of effects of wordlikeness in word recognition and production, providing evidence of the independent contribution of multiple lexical and form–based factors in the activation of a target word and its neighbours. In this section, we describe wordlikeness effects by relating them to specific aspects of TSOMs' behaviour.

First, Goldrick and colleagues observe a differential role of overall sublexical overlap (number of features/units shared by any two input strings) compared with position–specific overlap, resulting from a mapping between identical units taking the same relative position in two strings. Accordingly, target strings and their neighbours share position–specific unit representation (at least coarsely defined). We observe that, through recoding, TSOMs tend to enforce context–specific specialisation of map nodes: nodes that are sensitive to a particular symbol type are either strongly co–activated or localised in the same connected area of the map, with sub–clusters of nodes in the area responding to context–specific instances of the same symbol. Nonetheless, since context–specific specialisation is defined in terms of acquired sensitivity to the immediately preceding (left) context, nodes tend to reflect recurrent distributional and structural properties of the input language and are tolerant to time–shifts in sequential recoding (see GEMACHT–MACHT co–activation distances in Figure 4).

In addition to context–specific recoding of base units (segments or letters) a special status should be accorded to specific positions in the input string. In particular, letters in the initial position of words form a distinct group within lexical representations. TSOMs do not devote specific resources to symbols in the initial position of words to allow them to play a critical role in sequencing. Rather, they assume that lexical representations are retrieved sequentially from integrated (buffered) activation patterns, with activation spreading from left to right in hierarchically–arranged node graphs (Marzi et al. 2014). Due to this recall mechanism, nodes in initial positions establish a primacy gradient, so that they are less likely to be confused with other nodes further down in the input string and are recalled more accurately.

Left–to–right activation spreading through an integrated activation pattern is likely to activate irrelevant nodes (due to lexical competition) until the map reaches a point in the word graph where only one lexical alternative is available. This causes word endings to be retrieved and repeated more accurately than mid–word symbols, thereby producing a recency gradient. In addition, TSOMs can perceive the similarity between words ending in the same way (e.g. WRITING and WALKING, in our training set KOMMEN and NEHMEN), thus enforcing a structural alignment between sublexical chunks that may not be aligned in time (since they occur in words of different length). It is important to appreciate that perception of similarity at the outset of words plays a different role than perception of onset similarity, since the former is involved in inhibitory competition and the latter in lexical entrenchment vs. perception of structure and generalisation (Marzi & Ferro 2013).

Node buffering (through integrated activation patterns) is the primary cause of another well–known effect of word recall/repetition: longer words are recalled and repeated less easily than short words. TSOMs replicate this effect, since the number of nodes in the activation chain of a word may affect accuracy of serial recall for several reasons: in longer words node chains are more likely to exhibit weaker connections, due to a more combinatorial sequencing of both high– and low–frequency chunks; a larger number of nodes increases the chance of node confusability due to repetition of a symbol or a sequence of symbols; finally, primacy and recency effects make mid–word nodes more difficult to reinstate, causing a characteristically left–to–right U–shaped gradient of recall accuracy.

Word frequency is a hallmark of lexicality effects. High–frequency words are retrieved and repeated more easily and accurately than low–frequency words (Taft 1979; Alegre & Gordon 1999; Rastle et al. 2004; among others). Moreover, they appear to suffer less competition from neighbouring words and, in turn, to fiercely compete with low–frequency neighbours. TSOMs mimic word frequency effects through lexical entrenchment. First, connection weights between nodes responding to high–frequency words are proportional to the number of times a connection is traversed. This ensures that, other things being equal, a repeatedly–traversed node chain will show, on average, high connection weights, corresponding to stronger expectations for the word to be perceived. In addition, since presynaptic connections converging on the same node inhibit one another, high–frequency words tend to recruit "specialised" nodes, i.e. nodes that selectively respond to a single word only. Thus, entrenchment favours individual access and holistic perception, while disfavouring spreading of activation to other neighbouring forms and perception of internal structure. This dynamic has two consequences on word competition and word neighbourhood. If a low–frequency word shares the onset with a high–frequency word (e.g. the German inflected verb forms "GIBST–GIBT", the former being a hapax in our training set and the latter being highly frequent), there will be a stronger bias for the map to recall the high–frequency word when the low–frequency word is input (competition). On the other hand, if high–frequency words share a chunk after a different onset (e.g. MACHT and GEMACHT), we observe even activation levels and co–activation, reflecting a robust *priming* effect.

TSOMs thus make the prediction that accessing and recalling a word rely both on lexical entrenchment/expectation, and on global activation of a dense neighbourhood. We explore these effects in the following section.

5. The role of neighbourhood structure in word perception and production

Vitevitch et al. (1997) offered evidence of the facilitatory influence of phonotactic probabilities on processing time for spoken stimuli. They observed that mono– and bi–syllabic non–words composed of familiar segmental sequences of English were repeated faster than non–words composed of infrequent sequences, suggesting that facilitatory effects of probabilistic phonotactics reflect differences among activation levels of sublexical units (segments or sequences of segments).

This evidence appeared to contradict predictions made by the Neighbourhood Activation Model (NAM: Luce & Pisoni 1998) and other computational models of word recognition such as TRACE (McClelland & Elman 1986) and Shortlist (Norris, 1994). According to these models, words that are like many other words (e.g. words in a dense similarity neighbourhood) should be recognised more slowly and less accurately than words with fewer similar words, due to within–lexicon competition between concurrently activated words. Since high–density neighbourhoods demonstrably contain words with high–frequency n–grams, we are left with evidence of differential effects of phonotactic probabilities on processing words as opposed to non–words.

Vitevitch and Luce (1998, 1999) proposed to account for this seemingly paradoxical evidence by resorting to two explanatory mechanisms for word processing, hinging upon segregated (but interactive) representational levels for word encoding: a sublexical level, and a lexical level proper. Processing a novel input stream, irrespective of its being a word or a non–word, involves incremental activation of specialised nodes selectively attuned to respond to input symbols in context. At this sublexical level, expectations of up–coming symbols facilitate (FACILITATORY COMPETITION) and speed up recognition (and immediate recall). This is common to processing input non–words as well as input words. However, according to Vitevitch and Luce, when perception involves an actual stimulus word, activated sublexical nodes spread activation to the lexical level, bringing about identification of one word with its unique meaning and other properties at the stage of lexical access. It is at this point in time that a level of lexical processing sets in, with words in high–density lexical neighbourhoods mutually competing for primacy: the denser the lexical neighbourhood, the stronger the level of INHIBITORY COMPETITION. Inhibitory competition between simultaneously activated lexical nodes explains evidence of slowed down processing of words in dense neighbourhoods. Lexical competition damps any benefit these high–density words have from containing high–probability phonotactic patterns. Non–words do not suffer from the same problem, as they fail to activate word nodes on the lexical level.

TSOMs provide an interesting computational framework for implementing integrative models of lexical storage and processing, and can, in our view, shed some light on the interaction of inhibitory and facilitatory competition in lexical acquisition.

TSOMs do not define autonomous levels for form–based (either phonotactic or orthotactic) knowledge and lexical knowledge. Knowledge of sequential form–based probabilities is controlled by storage of lexical forms. This is reflected by the propensity of a TSOM to recode highly probable n–grams, i.e. n–grams frequently recurring in the lexical training set, as strongly connected node chains, whose level of activation is directly proportional to the strength of their connections. As a result, low–frequency n–grams are poorly recoded in the network, which shows little propensity for repeating them upon hearing, since each node in the chain has a low expectation for the ensuing node to be activated.

A non–word with high–frequency chunks is likely to overlap with many stored words at different positions in the string (corresponding to different n–grams.), thus being more likely to be repeated by the map as quickly and correctly as possible. Hence, non–words containing familiar chunks tend to be repeated faster (and more accurately) than non–word isolates. This is shown in Figure 10 (right), where we measured the average connection strength of correctly recalled vs. wrongly recalled node chains associated with non–words. The plot shows a statistically significant[5] advantage in connection strength of

[5] To verify hypothesis about the significance of differences in distribution, the T test (Student test) is used to compare possible population conditions by way of two competing hypotheses: the null hypothesis, which is the neutral or non–significant statement as no difference between distribution of two groups, and the research–hypothesis. The p–value indicates the probability of obtaining the observed sample results when the null hypothesis is actually true. For very small values of p–value (less than or equal to a threshold value – namely the significance level (traditionally 5% or 1%) – the observed data are inconsistent with the null hypothesis, that is the other hypothesis, the research hypothesis, is accepted as true.

correctly recalled non–words, bearing witness to their higher degree of lexical predictability, confirmed by the strong correlation between values of connection strength and orthotactic forward probabilities (r=0.812, p<.000005).

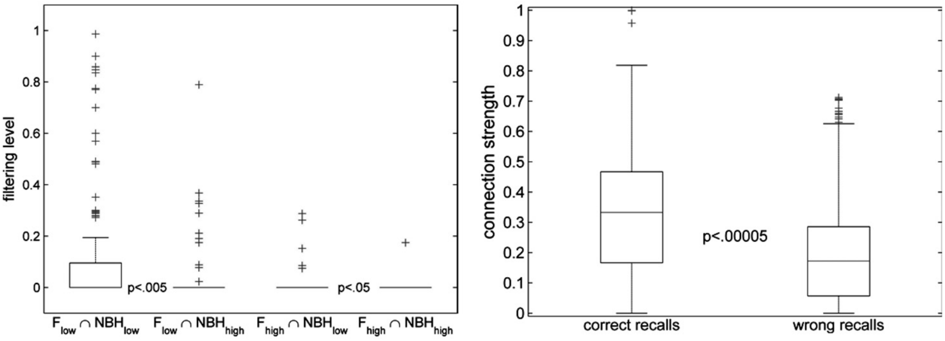


FIGURE 10: Filtering levels on word memory traces for serial recall (left). Low–frequency words in low–entropy neighbourhoods ($F_{low} \cap NBH_{low}$) are shown to require higher levels of filtering for them to be recalled correctly. Box plot distribution (right) of average connection strength of correctly recalled vs. wrongly recalled words. p–values give the significance level of the hypothesis of statistical independence between distributions.

But what about inhibitory effects in access/recall of input words with dense neighbourhoods? It should first be reminded that lexical processing and pre–lexical (or sublexical) processing interact dynamically in TSOMs. When the first symbol ’M’ of the input word “MACHEN” is shown to the map, it will partially activate all node chains starting with ’M’, with a strength that is proportional to the number of times M–starting words were shown to the map during training. Hence, while no single lexical entry has yet been fully accessed for its complete set of properties to be retrieved, several lexical entries receive “partial activation”. Likewise, when the symbol ’A’ is shown immediately after ’M’, the strength of expectation for the corresponding ’A’ node to be activated will be a direct function of the cumulative frequency of all stored words beginning with ’MA’. However, at this point in time, all words beginning with ’M’ but not with ’MA’ (e.g. MÖGEN, MÜSSEN etc.) will expect different continuation symbols than ’A’, thus pre–activating other competing nodes, and blurring the more prominent (and correct) activation chain with spurious activations. Thus partial lexical activation produces two effects. On the one hand, it provides a distinct benefit to the most typical n–grams attested in the lexicon, projecting a strong selective expectation on n–gram constituents, in terms of inter–node connections. These expectations represent the basis for n–gram probabilities. On the other hand, partial activation also implies that a number of alternative irrelevant sequences are being activated, thus creating noise that hinders lexical access/recall.

This behaviour was tested by simulating an immediate lexical recall protocol. First, the map was exposed to a target word, and then it was given the task of recalling the letters of the target word from its integrated activation pattern. We repeatedly checked how accurately the map could perform the task for increasing levels of activation filtering: starting from the noisiest condition (corresponding to an unfiltered integrated pattern, including all co–activated nodes), to arrive at a "skeletal" pattern, where only BMU nodes (with activation strength 1) are kept in the pattern. At each filtering step, we checked the map's recall, to eventually record the lowest level of filtering at which the map can reinstate the target word accurately. Since the integrated activation pattern of a word contains, for each word symbol, the concurrent activation of all possible nodes competing for that symbol, the level of filtering is an inverse function of how reliable the activation pattern is. A deeply entrenched node chain with few competitors is expected to be recalled with no filtering (level 0). Conversely a weak node chain, suffering from a pool of strong competitors, should require a high level of filtering to be recalled correctly. In the worst case, when the map was not able to recall the word from the activation pattern, the filtering level was set to 1.

The protocol is somewhat reminiscent of the immediate repetition task reported by Vitevitch et al. (1997). We then expected different levels of map's sensitivity to filtering, depending on both neighbourhood entropy and token frequency of the target word. In particular, low–frequency targets in a low–entropy neighbourhood should be the most difficult to recall, as they suffer from strong inhibitory competition and weak entrenchment. On the other hand, high–frequency targets in a high–entropy neighbourhood should be the easiest to recall.

The box plot in Figure 10 (left) confirms these expectations. It shows the distribution of levels of filtering for four classes of words, by combining frequency bins of target words (low vs. high) with entropy bins (low vs. high) of their neighbourhoods. Target words in the leftmost group on the *x*–axis correspond to the condition of most extreme inhibitory competition: a weak target surrounded by strong competitors. This condition consistently requires higher levels of filtering for the map to be able to recall the target word, by counteracting the pressure of aggressive competitors.

6. General discussion and concluding remarks

The relationship between perception of typicality and (morpho–)phonotactic probabilities and lexical density is strongly correlated with lexical organisation and processing. As observed by Bailey and Hahn (2001), sequence typicality of words – or wordlikeness – affects both language acquisition and processing. In this perspective, computer simulations of lexical storage provide a methodological framework for testing models of word processing and perception of typicality. In particular, in TSOMs acquisition strategies are analysed by focussing on emergent relations between stored word forms and on dynamic expectation/competition of incoming input. In good accord with a memory–based perspective, these strategies are highly dependent on input properties such as type and token frequency, as well as semantic and phonological consistency (Lieven & Tomasello, 2008; Tomasello, 2003).

In the unsupervised artificial neural network proposed here, the developmental course of word memory traces and their organisation and role in word perception, access and recall, can be simulated incrementally and monitored at a considerable level of detail. We can thus selectively explore the time–bound consequences of frequency and neighbourhood effects on the overall organisation of a lexical map and on its performance in well–defined tasks.

With these goals in mind, we tested a TSOM trained on German verbs by observing its behaviour in recalling words from three different test sets not included in training, and defining a natural gradient of word familiarity to a German TSOM: unknown forms of known verb paradigms (set ii), unknown German paradigms (set iii) and Italian paradigms (set iv).

To quantitatively assess the map's behaviour and correlate this behaviour with traditional mathematical models of wordlikeness (i.e. phonotactic/orthotactic probability and neighbourhood density), we observed the state of the map in recalling an input word, and recorded this state through three indices: (i) connection strength, (ii) activation strength, (iii) global co–activation strength.

Connection strength and activation strength appear to significantly correlate with each other, and are able to capture a notion of serial (or "syntagmatic") wordlikeness, based on the frequency distribution of recurrent (sublexical) chunks in the lexicon. Furthermore, they appear to characteristically define "word familiarity" as the degree of the map's expectation for a word form to occur. Serial wordlikeness is shown to play a fundamental role in both word recoding and recall, although its relative contribution is considerably more prominent in the latter task (where the strength of forward connections is critical for the map to reinstate a word form from its activation pattern), than in the former task (which is mostly signal–based). Accordingly, accuracy of word access and recall is demonstrably correlated with both connection strength and activation strength, which are, in turn, related to lexical frequency and lexical entrenchment effects.

Nonetheless, serial expectations define only one specific dimension of wordlikeness, which can be referred to as "first–order" or "syntagmatic" wordlikeness. Another complementary but correlated dimension is controlled by the notion of neighbourhood density. In TSOMs, neighbourhood density can be measured as the extent to which a particular input word reverberates in the global activation pattern of the acquired lexicon: the higher the reverberation, the more wordlike a form is perceived, and the larger the number of its neighbours. Unlike "syntagmatic" wordlikeness, this notion has little to do with local perceptual salience, and does not suffer from effects of time–alignment and lexicality. Rather, it has to do with global lexical coherence, by saying how well a word fit into the global organisation of the lexicon. A high value on this dimension means that the overall structure of the input word form is familiar, irrespectively of whether it is more or less difficult to recall. We may refer to this dimension as "second–order" or "paradigmatic" wordlikeness, since it probes the overall organisation of the lexicon and the degree of mutual relationship among stored words (rather than the syntagmatic well–formedness of word internal constituents). Second–order wordlikeness presupposes first–order wordlikeness.

This is shown in Figure 7, where novel German words from both known and unknown paradigms exhibit comparatively high levels of global co–activation, in contrast with Italian words, whose degree of "paradigmatic" wordlikeness is consistently poorer (p<.00001). We can explain this overall effect of German–likeness by looking at differential values of activation strengths for stems and suffixes in Figure 8. Here, unknown German words with known inflections (word set iv) score more highly on suffixes than on stems. As expected, they are recalled consistently more poorly (Table 1), but their degree of perceived familiarity can be accounted for by their fitting a recurrent morphological pattern exhibited by other German words the TSOM is already familiar with. Clearly, this does not apply to Italian word forms, which can occasionally exhibit German–like chunks, but comply more hardly with German morphological patterns.

Second–order wordlikeness thus defines a type of non–serial constraints that go beyond n–gram probabilities, and can be described in terms of structure–based familiarity. These constraints regulate the density of a word's neighbourhood, with words scoring high on second order wordlikeness being surrounded by more neighbours.

High activation values for misrecalled novel German forms reflect these constraints, with misrecalled Italian forms being, whenever possible, forced to comply with them. Given the specific nature of our training data set (a sample of the most frequent German verb forms), the map's sensitivity to structure eventually reflects morphological constraints holding across the entire German verb system. Marzi and colleagues (2014) provide considerable computational evidence of interactive effects of morphological regularity and type and token frequency in the acquisition of German verb inflection.

The wordlikeness effects reported here point in the same direction, suggesting that a unitary model of lexical memory can explain both types of evidence under the same set of principles. In a nontrivial sense, inhibitory effects in low–entropy neighbourhoods and, conversely, facilitatory effects in high–entropy neighbourhoods are accountable in terms of the same co–activation/competition dynamics governing evidence of easier accessibility of words in highly entropic word families, as reported by Moscoso and colleagues (2004). Having established a deep interconnection between frequency–based effects of word similarity and lexical redundancy on the one hand, and principles of input recognition and recall on the other hand, is a non marginal result of the experimental evidence offered here.

References

Aitchison, J. (1987). Words in the Mind: An Introduction to the Mental Lexicon. Oxford, Blackwell Publishers.

Alegre, M. & P. Gordon (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40: 41–61.

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

Bailey, T. M., & U. Hahn (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44.4: 568–591.

Bard, E. G. (1990). Competition, lateral inhibition, and frequency: Comments on the chapters of Frauenfelder and Peeters, Marslen–Wilson, and others. In G. T. M. Altmann (ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. Cambridge: MIT Press. 185–210.

Bybee, J. (1995). Regular Morphology and the Lexicon. *Journal of Verbal Learning and Verbal Behavior* 10.5: 425–455.

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.

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. Bologna: Il Mulino.

Goldrick, M, J. R. Folk & B. Rapp (2010). Mrs. Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language* 62.2: 113–134.

Gupta, P. (2005). Primacy and recency in nonword repetition. *Memory* 13.3–4: 318–324.

Gupta, P. (2009). A computational model of nonword repetition, immediate serial recall, and nonword learning. In A. Thorn and M. Page (eds.), *Interactions between short–term and long–term memory in the verbal domain*. Hove, UK: Psychology Press.

Hahn, U., & T. M. Bailey (2005). What makes words sound similar? *Cognition* 97.3: 227–267.

Hickok, G. & D. Poeppel (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92.1: 67–99.

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

Lieven, E. & M. Tomasello (2008). Children's first language acquisition from a usage–based perspective. In P. Robinson & N. Ellis (eds), *Handbook of cognitive linguistics and second language acquisition*. New York: Routledge. 216–236.

Luce, P. & D. Pisoni (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and hearing* 19.1: 1–36.

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

Marzi C., M. Ferro & V. Pirrelli. (2012a). Prediction and Generalisation in Word Processing and Storage. In *8th Mediterranean Morphology Meeting Proceedings on Morphology and the architecture of the grammar*. 113–130.

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

Marzi C. & M. Ferro (2013). Adaptive strategies in lexical acquisition. *Lingue e linguaggio*, 12.2: 307–328.

Marzi C., M. Ferro & V. Pirrelli (2014 forthcoming). Morphological structure through lexical parsability. *Lingue e linguaggio* 13.2.

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

Mirman, D., T. J. Strauss, J. A. Dixon & J. S. Magnuson (2010). Effect of representational distance between meanings on recognition of ambiguous spoken words. *Cognitive Science* 34.1: 161–173.

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

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52: 189–234.

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

Rastle, K., Davis, M. H., New (2004). The broth in my brother's brothel: Morpho–orthographic segmentation in visual word recognition. *Psychonomic Bulletin and Review* 11.6: 1090–1098.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory and Cognition* 7: 263–272.

Tomasello, M. (2003). Constructing a language: A usage–based theory of language acquisition. Cambridge, MA: Harvard University Press.

Vitevitch, M. S., P. A. Luce, J. Charles–Luce, D. Kemmerer (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and speech* 40.1: 47–62.

Vitevitch, M. S. & P. A. Luce (1998). When words compete: Levels of processing in perception of spoken words. *Psychological science* 9.4: 325–329.

Vitevitch, M. S. & P. A. Luce (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40.3: 374–408.

Vitevitch, M. S., J. Armbrüster & S. Chu (2004). Sublexical and lexical representations in speech production: effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30.2: 514–529.

Percepcija tipičnosti u leksikonu: tipičnost oblika riječi, leksička gustoća i morfonotaktička ograničenja

Pokazano je da stupanj do kojeg je određeni simbolički vremenski slijed (slijed zvukova ili slova) tipična riječ u jeziku, odnosno TIPIČNOST OBLIKA RIJEČI, ima učinaka u proizvodnji i percepciji govora, uspješnosti čitanja, leksičkom razvoju i pristupu leksemima te kratkotrajnoj i dugotrajnoj verbalnoj memoriji. Predložena su dva kvantitativna modela kako bi se objasnili navedeni učinci: serijalne fonotaktičke vjerojatnosti (vjerojatnost pojavljivanja određenog simboličkog slijeda u leksikonu) i leksička gustoća (mjera do koje se druge riječi mogu proizvesti zamjenom, brisanjem ili umetanjem jednog ili više simbola u ciljnu riječ). Te dvije mjere visoko koreliraju, zbog čega su teško razdvojive pri mjerenju njihovih učinaka u leksičkim zadacima. U ovom radu predlažemo računalni model leksičke organizacije koji pruža sustavan algoritamski prikaz učinaka leksičkog usvajanja, obrade i pristupa kako bi se dodatno istražila ova pitanja. Taj se model temelji na samoorganizirajućim mapama s hebijanskim vezama definiranim preko vremenske razine (engl. TSOMs). Posebice pokazujemo da se (morfo-)fonotaktičke vjerojatnosti i leksička gustoća, iako korelirani u leksičkoj organizaciji, mogu shvatiti kao načini usredotočavanja na različite aspekte govornikova ponašanja pri obradi riječi i tako pružiti nezavisne kognitivne doprinose našem razumijevanju principa percepcije i tipičnosti koji upravljaju leksičkom organizacijom.

Key words: lexical acquisition, word processing, frequency, mental lexicon
Ključne riječi: usvajanje leksika, procesiranje riječi, frekvencija, mentalni leksikon