A Recurrent Multimodal Network for Binding Written Words and Sensory-Based Semantics into Concepts

Andrew P. Papliński[1], Lennart Gustafsson[2], and William M. Mount[1]

[1] Monash University, Australia
Andrew.Paplinski@monash.edu
[2] Luleå University of Technology, Sweden
Lennart.Gustafsson@ltu.se

Abstract. We present a recurrent multimodal model of binding written words to mental objects and investigate the capability of the network in reading misspelt but categorically related words. Our model consists of three mutually interconnected association modules which store mental objects, represent their written names and bind these together to form mental concepts. A feedback gain controlling top-down influence is incorporated into the model architecture and it is shown that correct settings for this during map formation and simulated reading experiments is necessary for correct interpretation and semantic binding of the written words.

Keywords: Words and Concepts, Multimodal binding, Self-Organizing networks, Bigrams.

1 Introduction

We use a network of recurrent self-organizing modules to model aspects of the reading process within the cortex. As perceptions of the world around us are experienced by combining sensory inputs of different modalities with internal world-models learned within the mind, our network consists of models of five cortical areas, two of which process the sensory information and three others represent a two-level hierarchical model of the world. Such an architecture is motivated by the fact that the neural processing first takes place in mainly unimodal (visual, auditory, etc.) hierarchies in the brainstem and sensory cortices of the cerebrum. The unimodal percepts then converge in multimodal association areas such as STS (Superior Temporal Cortex). At this level we have highly abstracted, semantic representations of objects. We attach words to these mental objects and thus build conceptual representations.

The world in our work consists of a set of animals defined in terms of perceptual features and qualities and we bind the written animal names to the learned mental objects representing them. The processing and binding of written names to mental objects follows a similar methodology to [19,9]. However, we have improved working of the network by adding feedback gains to control the level of modulation feedback from the modeled bimodal integration area.

B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part I, LNCS 7062, pp. 413–422, 2011.

The neuronal circuitry involved in reading is undoubtedly complex and much current research concentrates on an area called VWFA (Visual Word Form Area) in left fusiform gyrus [13] where prelexical, i.e., strings of letters, and lexical processing of word forms [5] takes place. One of the more complete representations of cortical areas involved in the process of reading resulting from intensive fMRI investigation is given in [4]. Fig.2.1 (available electronically) presents 13 interconnected cortical areas, arranged in five groups: visual input, visual word form, access to meaning, access to pronunciation and articulation, and top-down attention and serial reading. In our work we use a much simplified model consisting of just five 'cortical' areas. One of the basic premises of our modelling framework is the concept of a ubiquitous "neuronal code", which implies a unified way of representing information exchanged by modules of the network.

At the beginning of our consideration is the problem of coding words in neocortex, since several methods based on letter combinations and positioning with increasing sophistication have been proposed [3,8,20]. We will employ here a relatively straightforward method called open bigram coding as presented in [20,4] and described further in Section 3.1. The other fundamental consideration is how and where conceptual representations are stored in neocortex. We adopt the unitary system hypothesis as argued in [1].

2 Model Description

As a hierarchical model of reading combining bottom-up sensory integration with top-down processing our model follows principles similar to those presented in [7], which describes a multi-layered model for processing of features, letters and words in cortex. For a related neurocomputational account of map-based processing and its role in language and speech comprehension and production, see also [11,12].

In our approach, which stems from our earlier works on multimodal integration [19,2,18] self-organized modules form low dimensional labels. These labels are used as afferent signals to up-stream modules, and may represent any type of perceptual, conceptual or lexical 'features'. Such universal feature labelling is consistent with the neurocomputational modelling approach of [6].

As noted above, we do not attempt to represent each and every cortical area taking part in perceptual and semantic processing of mental objects or in higher-level language and related cognitive tasks. Rather our aim is to represent a subset of language processing in a smaller number of modules aggregating the processes of several cortical areas. The model consists of 5 processing modules divided into 3 layers, connected as shown in Figure 1. The processing pathway for visual and associated perceptual information for the mental object is depicted on the left hand side of the figure (P and MO) while word processing paths are included on the right (Wrd and UW). These pathways then converge and binding occurs within a hypothetical high-level bimodal association area, M+W.

Each module consists of a number of artificial neuronal units randomly located in a circular area. Relative positions of 'neurons' inside the circle are described by a position matrix **V**. The total number of neurons is selected in proportion

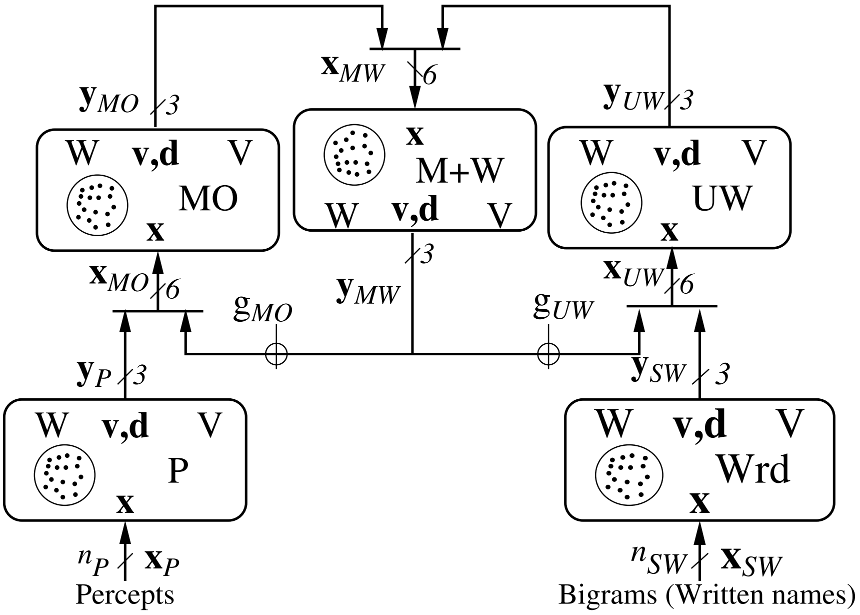

Fig. 1. The network of simulated cortical maps: Wrd – Sensory Word map, UW – Unimodal Association Word map, P – Sensory Percepts map, MO – Mental Objects map, M+W – Bimodal Association map: mental objects and written names

to the number of objects represented by the module, with rows in the weight matrices **W** characterizing the synaptic strengths in each module. The main functions for each map is described below:

- Percepts map (P) encodes basic perceptual features, such as size, colour, form of locomotion and social behaviour for each object within the category: *Animals*.
- Mental Objects map (MO) is a map of perceivable "mental objects" and semantic relationships for the object category, i.e. a topographically organised map of a set of 30 animals arranged according to perceptual features.
- Word map (Wrd) models letter and letter position processing in VWFA. Input to this module take the form of word bigrams based on the possible pairings of 26 English/Latin letters.
- Unimodal Word map (UW) performs sub-lexical processing of words and projects these as written names for lexical binding within the bimodal association map.
- 'Bimodal' association map (M+W) is responsible for binding perceivable mental objects to their written names in order to form a set of labelled mental concepts.

Sensory modules in Figure 1 operate on a relatively large number of afferent signals and produce a low-dimensionality efferent signal labeling the sensory object. Association modules, in turn use these labels as their afferent signals and produce efferent labels encoding positions of activated neuronal patches within these maps. The operation of a cortical module is functionally equivalent to mapping the higher dimensionality input space to a three-dimensional output space. More specifically, the label information consists of a three-dimensional vector encoding the relative location of the most excited neuron within a given cortical patch, supplemented by the post-synaptic activity level of this neuron.

3 Operation of Recurrent Network

The processing modules within the network model are interconnected via feed-forward sensory processing pathways and feedback connections from the bimodal integration module (M+W) (see Figure 1). Feedback gain terms, g_{mo} and g_{uw} are also shown on these recurrent pathways. The effect of these is discussed below and in Section 3.2.

It is assumed that during normal operation association modules perform static non-linear mappings of the form:

$$\begin{aligned}
\mathbf{y_{MW}} &= f(\mathbf{W_{MW}} \cdot \mathbf{x_{MW}}) \; , \;\; \mathbf{x_{MW}} = [\mathbf{y_{MO}}, \mathbf{y_{UW}}] \\
\mathbf{y_{MO}} &= f(g_{MO} \cdot \mathbf{W_{MO_{MW}}} \cdot \mathbf{y_{MW}} + \mathbf{W_{MO_P}} \cdot \mathbf{y_P}) \\
\mathbf{y_{UW}} &= f(g_{UW} \cdot \mathbf{W_{UW_{UW}}} \cdot \mathbf{y_{MW}} + \mathbf{W_{UW_{SW}}} \cdot \mathbf{y_{SW}})
\end{aligned}$$

where $\mathbf{x}$ and $\mathbf{y}$ represent input and output signals, respectively and $f(\cdot)$ describes the Winner-Takes-All function. For the bimodal association module we can write the following dynamic equation:

$$\mathbf{y_{MW}}(t+1) = \mathcal{M}(\mathbf{y_{MW}}(t), \mathbf{y_{MO}}(t), \mathbf{y_{UW}}(t)) \tag{1}$$

which formally describe the recurrent and non-linear nature of the network that may result in complex time behaviour. From the simulation perspective, the trained network is observed to settle immediately if we apply exogenous stimuli that are congruent with the endogenous thoughts or initial conditions. In other words, if the labels in the network are known and congruent, the network quickly converges to a stable state. See [19] for behaviour of a similar network for incongruent inputs.

3.1 Preprocessing of Percepts and Word Bigrams

Prior to training of the maps, a preprocessing step is performed to produce the sensory-based semantic and letter bigram information for the separate perceptual and written word/lexical pathways. Open bigram encoding [20,4] is used in the present model. The purpose of this is to encode attribute lists of the former and relative letter positioning of associated words for the later into a consistent numerical format for the self-organised maps. To ensure computational efficiency all inputs and weight vectors are projected on the unity hypersphere. Working with unity vectors makes it easy to compare them by calculating relevant inner products and allows us to use a simplified dot-product learning law [10].

3.2 Sequential Development of Maps

The map training sequence approximately follows that of the widely accepted model of neural ontogenesis and cortical map formation. In general, maps and their interconnections are trained using the Kohonen learning law[10], normalised Hebbian learning [14,15,16], or combination of both as in [11,12]. In particular, the processing and adaptation to sensory and learned label information is propagated in a feedforward or bottom-up direction and feedback processing subsequently comes into play in a recurrent optimisation of the higher level maps.[1]

Feedforward training of unimodal sensory maps. The first training step involves initial organisation of the unimodal maps, MO and UW. To train the map of mental objects, MO, feature vectors $\mathbf{x_P}$ describing the animals are first encoded by the auxiliary module, P as a dimensionally reduced label, $\mathbf{y_P}$ and used as input to MO. A competitive learning process then encodes these inputs as a map of mental objects organised according to their perceptual semantics. Following training, the object categorizing module P is disconnected and the signal $\mathbf{y_P}$ is interpreted as *thought commands* used for recalling items stored within module MO. The UW map is trained independently using the topographically organised word bigram representations from the Wrd module.

Feedforward training of bimodal integration map. The next step is to train the bimodal map M+W using combined inputs from MO and UW. Through statistical pairing of randomly presented mental object and word label information from each of the unimodal maps, an initial bimodal map of lexically-bound *concepts*, in this case of a named set of animals, is formed.

Feedback training of sensory and bimodal maps. Following completion of the feedforward training steps, the three association modules forming the recurrent part of the network are trained together. This time, after each learning step we perform several relaxation steps running the network as in Eqn (1), until all efferent signals in the network are constant. There is a limit on the number of iterations imposed that is important in the initial learning stages. Once the maps are fully developed, the network stabilizes quickly after a small number of iterations depending on congruence between the perceived mental objects and associated written words or names.

As a result of this training step, the unimodal maps, MO and UW are optimised and re-organised to reflect contextual information transmitted from the bimodal integration layer. The M+W module in turn is adjusted so as to represent the statistical correlations between the unimodal data encoded in the label information $\mathbf{y_{MO}}$ and $\mathbf{y_{UW}}$. A notable new feature of the model architecture presented in Figure 1 are the feedback gain terms, g_{MO} and g_{UW}. The effect of reducing the feedback gain is to attenuate the significance of the feedback relative to feed-forward signals.

[1] Note that while no recurrent feedback from higher level modules is used in the initial feedforward training steps, self-organisation of the maps assumes local recurrent connections across the map output layer in order to implement the required competitive learning process.

In order to ensure that inputs to the MO and UW modules are properly bounded and that weight vectors remain on the unity hypersphere, a novel algorithm is employed in which the inputs and gain values are effectively spatially decomposed and then re-assembled within the neuronal units.

3.3 Testing Response of Network to Word Stimuli

Following training of the maps, the operation of the network is essentially as follows: The written word excites cortical patches in the word bigram map, Wrd, and via the forward path from this low-level sensory area, patches within the unimodal association area UW and bimodal binding area, M+W are also excited. A similar pattern of activation is produced via the feed-forward pathway from the 'perceptual' mental objects area MO to M+W. These direct feed-forward paths assure that the binding process is rapid – at least in the case of congruous thoughts and inputs. In the case of discrepancy or incongruity between the mental object and a written word (for example misspelt word, or mismatch between word and perceived object) a feedback loop is automatically activated from the M+W area back to the MO and UW modules. This initiates a recurrent cycle that typically converges on a globally sensible solution, assuming that no contradictory input is presented and that one actually exists.

As in the feedback training step, the gain terms g_{MO} and g_{UW} can be set separately during the testing or operational phase. In this way the effect of controllable feedback upon the hierarchically organised set of trained maps, or equivalently, top-down influence on the simulated reading process can be explored. An initial set of results showing some of these effects are presented below.

4 Simulation Results

Preliminary results show the development of activities over time in the modules of our simulated network when thought commands about different mental objects, in this case animals and their associated written names are simultaneously presented. This perceptual or mental binding task is considered central to language understanding during the activity of reading. The simulation scenario is also similar to the fMRI-based study of recognition of spoken words representing animals when subjects are cross-modally primed for different animals [17].

As an example of the operation of the network in recognition of words corresponding to known mental objects, consider the situation where the multimodal computational network is initialised to a particular animal, while being presented with the perceptual features of another animal and a misspelt version of the word for that animal. The response of the network to this situation is presented in Figures 2 and 3.

In the maps depicted in these figures, neuronal positions are marked with the yellow dots and the map area is tessellated with respect to the peaks of neuronal activities for each stimulus. We can identify the neuronal patches for various animals, e.g., 'dog', 'frog' and 'goat'. The objects or written names are 'placed' in the relevant cortical area during the learning process. Once the learning is

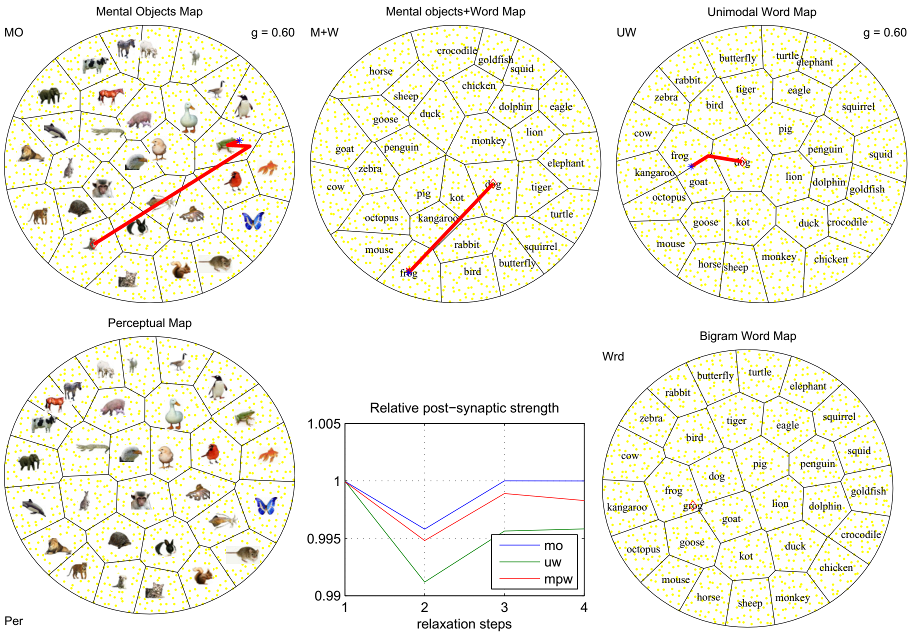

Fig. 2. The relaxation trajectories in the association maps when the word map, Wrd, presents a misspelled word "grog", the perceptual map, Per, presents object "frog", and the initial state was "dog". Both feedback gains are as during training, (g=0.6).

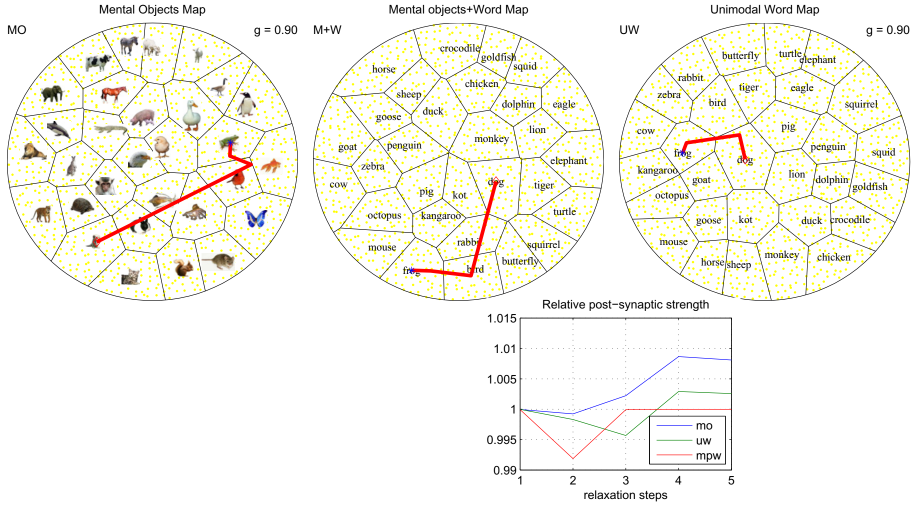

Fig. 3. The same "frog–grog–dog" trajectories for the high feedback gains, g=0.9

completed, the cortical area responds with an activity pattern characteristic to each stimulus.

At the starting point, the MO module *thinks* 'dog', whereas the unimodal word association module UW *reads* the orthographically similar word 'grog'. The bimodal association map, M+W is arbitrarily initialized with a mental object 'dog'. This initial state is marked on three maps in Figures 2 and 3 with the '$\diamond$'. Since these positions have been learned during the training procedure, the relative post-synaptic strengths are initially at their maximum value, as seen in the upper right hand side of both figures.

In the first case, Figure 2, relatively low feedback gains, $g_{UW} = g_{MO} = 0.6$ have been used during the reading test phase. In this scenario, although the bimodal map M+W and mental objects map, MO converge to a 'frog' solution, the response of the unimodal word area UW arrives at a point on the class boundary between 'frog' and 'goat'. This indicates uncertainty in recognition of the word for the animal most closely matching the stimulus ('grog'), implying that the network as a whole has not been able to successfully bind the correct name 'frog' to the corresponding mental object. This level of uncertainty is indicated by the final value of relative post-synaptic strengths, in which the response of MO and M+W is stronger than that of UW.

Now compare this with the situation in Figure 3 when a greater level of feedback gain, $g_{UW} = g_{MO} = 0.9$ has been used during testing. In this case the network quickly settles to a solution when the thought prevails; the final response of all maps including UW is consistent and the network as whole converges rapidly on the 'frog' conclusion. The level of confidence in this outcome is further indicated by the high levels of the relative post-synaptic strengths.

Effectively, the network has acted to correct the spelling of the distractor word through application of contextual knowledge contained within the interconnected association modules. This result is a simple demonstration of how *semantic priming* on a known set of mental objects within a category can be effectively represented within the model. In probabilistic (empirical Bayes) terms, it also suggests how new beliefs, hypotheses or perceptions of the world can be inferred when a network conditioned by given prior beliefs or initial conditions is presented with new sensory evidence.

5 Discussion

For the direct comparison above to be possible, it is necessary that exactly the same trained hierarchical network be used in each case. In this example, a feedback gain setting of $g_{UW} = g_{MO} = 0.6$ was used during recurrent phase in the formation process (as described in Section 3.2). More varied simulation results can be obtained if different feedback gain values are used, however due to space limitations examples of the types of aberrant behaviours that can be produced as a result are not considered at this time.

In general, reducing the feedback gain during map formation will result in a overall network that responds well to new and less predictable inputs (such as words and non-words) but which lacks the contextual knowledge required to correctly associate these inputs with cross-modal percepts and mental objects. Conversely, applying too great a level of feedback during recurrent training step results in a network with a tendency to become "locked up" in previously known states or *thoughts* and less able to adapt to new sensory information.

This suggests that an optimal level of feedback gain is required in order to realise a reading network which can effectively employ previously learned knowledge to correctly perceive and learn new words.

One possibility for future research would be to use the feedback gain within an incremental learning regime in which global reinforcement feedback is used to assess the utility of the learned set of maps at a particular setting of feedback versus feedforward bias. The feedback gain g could then be decreased if the network became 'stuck' and unable to adapt to new inputs or sensory evidence and increased if a stronger belief in the prior state or conditions was deemed to result in a better overall performance. Adopting of such a 'self-supervised' approach could be a way to incorporate a process analogous to *selective attention* in a straightforward and integrated way which works to optimise the efficiency of the learning process.

The complete simulation software is available from the first author.

6 Conclusion

We have presented a model for binding written names to perceptually-based semantic objects and provide preliminary results to demonstrate how this can be used in modelling cognitive functions basic to reading. This includes automatic 'correction' of misspelt words when a similar known word that is bound to an active mental object or by extension, object category is attended to. Such cognitive processes are of fundamental importance to the particular human activity of reading. The introduction of controllable feedback gain increases the behavioural repertoire of the model, presenting an opportunity to explore a number of other effects on learning and cognitive behaviour within the outlined computational framework.

In the interest of maintaining structural simplicity, several assumptions have been made in the model. For convenience specific visual and other perceptual modalities, conceptual categories and semantic relationships are combined in a "mental objects map". This simplification is computationally efficient as it allows this information to be encoded as arbitrary lists of attributes. In a more comprehensive and biologically realistic model, the auxiliary 'P' module could be divided into specific sensory modalities or sub-modalities, used to represent visual, tactile, spatial or other features related to mental object categories such as plants, animals or tools.

The representation of the mental objects would then involve a multimodal integration of such features and binding of these to their associated names. From the lessons gained through this modelling exercise and through experiments in lexical binding of spoken names to mental objects, we hope to extend the model to integrate structurally separate processing pathways and perform multimodal and *transmodal* binding across auditory and written language modalities.

References

1. Bright, P., Moss, H.E., Tyler, L.K.: Unitary vs multiple semantics: PET studies of word and picture processing. Cerebral Cortex 89, 417–432 (2004)
2. Chou, S., Papliński, A.P., Gustafsson, L.: Speaker-dependent bimodal integration of Chinese phonemes and letters using multimodal self-organizing networks. In: Proc. Int. Joint Conf. Neural Networks, Orlando, Florida (August 2007)
3. Davis, C.J., Bowers, J.S.: Contrasting five different theories of letter position coding: Evidence from orthographic similarity effects. J. Exp. Psych.: Human Perception and Performance 32(3), 535–557 (2006)
4. Dehaene, S.: Reading in the Brain. Viking (2009), `http://pagesperso-orange.fr/readinginthebrain/figures.htm`
5. Glezer, L.S., Jiang, X., Riesenhuber, M.: Evidence for highly selective neuronal tuning to whole words in the "Visual Word Form Area". Neuron 62(2), 199–204 (2009)
6. Gliozzi, V., Mayor, J., Hu, J.F., Plunkett, K.: Labels as features (not names) for infant categorisation: A neuro-computational approach. Cog. Sci. 33(3), 709–738 (2009)
7. Graboi, D., Lisman, J.: Recognition by top-down and bottom-up processing in cortex: The control of selective attention. J. Neurophysiol. 90, 798–810 (2003)
8. Grainger, J.: Cracking the orthographic code: An introduction. Language and Cogn. Processes 23(1), 1–35 (2007)
9. Jantvik, T., Gustafsson, L., Papliński, A.P.: A self-organized artificial neural network architecture for sensory integration with applications to letter–phoneme integration. Neural Computation, 1–39 (2011), doi:10.1162/NECO_a_00149
10. Kohonen, T.: Self-Organising Maps, 3rd edn. Springer, Berlin (2001)
11. Li, P., Zhao, X., MacWhinney, B.: Dynamic self-organization and early lexical development in children. Neuron 31, 581–612 (2007)
12. Mayor, J., Plunkett, K.: A neurocomputational account of taxonomic responding and fast mapping in early word learning. Psychol. Rev. 117(1), 1–31 (2010)
13. McCandliss, B.D., Cohen, L., Dehaene, S.: The visual word form area: expertise for reading in the fusiform gyrus. TRENDS Cog. Sci. 7(7), 293–299 (2003)
14. Miikkulainen, R.: Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. Brain and Language, 334–366 (1997), `http://nn.cs.utexas.edu/miikkulainen:bl97`
15. Miikkulainen, R., Kiran, S.: Modeling the Bilingual Lexicon of an Individual Subject. In: Príncipe, J.C., Miikkulainen, R. (eds.) WSOM 2009. LNCS, vol. 5629, pp. 191–199. Springer, Heidelberg (2009)
16. Monner, D., Reggia, J.A.: An unsupervised learning method for representing simple sentences. In: Proc. Int. Joint Conf. Neural Net., Atlanta, USA, pp. 2133–2140 (June 2009)
17. Noppeney, U., Josephs, O., Hocking, J., Price, C., Friston, K.: The effect of prior visual information on recognition of speech and sounds. Cerebral Cortex 18, 598–609 (2008)
18. Papliński, A.P., Gustafsson, L.: Feedback in Multimodal Self-organizing Networks Enhances Perception of Corrupted Stimuli. In: Sattar, A., Kang, B.-h. (eds.) AI 2006. LNCS (LNAI), vol. 4304, pp. 19–28. Springer, Heidelberg (2006)
19. Papliński, A.P., Gustafsson, L., Mount, W.M.: A model of binding concepts to spoken names. Aust. Journal of Intelligent Information Processing Systems 11(2), 1–5 (2010)
20. Whitney, C.: Comparison of the SERIOL and SOLAR theories of letter-position encoding. Brain and Language 107, 170–178 (2008)