Constraints on generalisation in a self-organising model of early word learning

MAYOR, Julien, PLUNKETT, Kim

Abstract

We investigate from a modelling perspective how lexical structure can be grounded in the underlying speech and visual categories that infants have already acquired. We demonstrate that the formation of well-structured categories is an important prerequisite for successful generalisation of cross-modal associations such that even after a single presentation of a word-object pair, the model is able to generalise to other members of the category. This ability to generalise a label to objects of like kinds, commonly referred to as the taxonomic assumption, is an emergent property of the model and provides an explanatory framework for understanding aspects of infant word learning. Furthermore, we investigate the impact of constraints imposed on the Hebbian associations in the cross-modal training phase and identify the conditions under which generalisation does not take place.

Reference


Available at:
http://archive-ouverte.unige.ch/unige:22701

Disclaimer: layout of this document may differ from the published version.
Constraints on generalisation in a self-organising model of early word learning

Julien Mayor* and Kim Plunkett

Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, OX1 3UD, United Kingdom
*E-mail: julien.mayor@psy.ox.ac.uk

We investigate from a modelling perspective how lexical structure can be grounded in the underlying speech and visual categories that infants have already acquired. We demonstrate that the formation of well-structured categories is an important prerequisite for successful generalisation of cross-modal associations such that even after a single presentation of a word-object pair, the model is able to generalise to other members of the category. This ability to generalise a label to objects of like kinds, commonly referred to as the taxonomic assumption, is an emergent property of the model and provides an explanatory framework for understanding aspects of infant word learning. Furthermore, we investigate the impact of constraints imposed on the Hebbian associations in the cross-modal training phase and identify the conditions under which generalisation does not take place.

1. Introduction

A central issue in early lexical development is how infants constrain the possible meanings of words to refer to objects of like kind. It is often assumed when infants learn a label for an object, they can apply it to the whole category. The generalisation of labels to new instances of objects within the same category is often referred to as the taxonomic assumption.1 Many researchers have suggested that babies make use of this taxonomic assumption, along with a series of other constraints, in order to narrow the hypothesis space when learning new word-object associations.2–4

Markman proposed that even though infants find thematic relations between objects salient and interesting (e.g. dog and bone), a taxonomic assumption is used when generalising labels,1 thereby overriding the thematic association. A series of studies have proposed that the taxonomic constraint is an evolved version of perceptually-based categorisation.5,6 A notable example of which is known as the shape bias,3,7 where infants show
a clear preference to group items according to their overall shape.

Although there is a wealth of empirical data characterising these phenomena, very little is known about the nature of the neural mechanisms underlying them. We propose that such constraints are emergent properties of the underlying neural architecture. We use a model made out of two Self-Organising Maps connected together with associative links. Self-organising maps (often referred to as SOMs\(^8\)) are good candidates for modelling the underlying mechanisms responsible for forming categories out of a complex input space; they achieve dimensionality reduction and auto-organisation around topological maps.\(^9\)

Previous studies have highlighted the promising role of SOMs as models of early lexical development.\(^10,11\) However, most of these studies have used heavily pre-processed input representations. In contrast we apply SOMs to \textit{real} auditory and visual input using Hebbian learning to form cross-modal associations and examine in a computational framework how categorisation and generalisation can emerge. We will report results on the ability the network has to generalise word-object associations and discuss the implications for the taxonomic assumption.

We ran three experiments in order to investigate the constraints on the generalisation properties of the network. In experiment 1, we assess the role the number of word-object pairings plays in generalisation. We demonstrate that even following a single word-object pair presentation, the network is able to generalise the label to other images of like kind and vice-versa. Even though classification success increases along with the number of trained word-object pairs, the normalised generalisation is approximately independent of the number of pairings. This suggests that generalisation properties depend on the \textit{physical organisation} of the model. Hence, we ran a second experiment in order to assess the influence of the map structure quality on generalisation capacity. We show that well structured maps are a prerequisite for good generalisation. Finally in experiment 3, we show how generalisation is limited by the number of units that are connected through Hebbian associations. Moreover, we show that generalisation performance reaches a peak even when only a limited number of units are allowed to \textit{fire and wire} together, satisfying a constraint of limited synaptic resources.

2. Method

Our model consists of two SOMs each receiving input from one modality, either visual or acoustic. In a first phase of training, maps are \textit{independently} fed with their respective input so that structure emerges. This first phase
models the early experience of a baby discovering the environment, by sampling her visual and acoustic surroundings. In a second phase of training, we connect both maps with associative links. This second phase captures the increasing importance of shared attentional activities, such as gaze sharing, joint attention or pointing at objects, during later infancy. Although the integration of both maps probably occurs gradually in the real world, for the sake of simplicity we wait for the maps to be structured and then build associations. Through *simultaneous* presentation of both a visual token and an acoustic token that belong to the same category (e.g., mimicking the behaviour of a caregiver pointing at a dog while saying the word “dog”), synapses connecting active nodes on both maps are reinforced. After Hebbian associations between the two maps are formed, we test the model by presenting an image and measuring the activity patterns produced on the auditory map (or vice-versa). Successful generalisation occurs if unpaired images from the same visual category produce activation on the auditory map corresponding to tokens of the same label.

2.1. **Training the Unimodal Maps**

The algorithm of self-organisation is the standard Kohonen algorithm.\(^8\) Each map (acoustic and visual) consists of an hexagonal grid of neurones receiving acoustic and visual inputs, respectively. With each neurone \(k\) is associated a vector \(m_k\). For the presentation of each input pattern \(x\), the vectors \(m_k\) are modified according to the following procedure:

We find the Best Matching Unit (BMU) \(i\), defined by the condition

\[ ||m_i - x|| \leq ||m_j - x|| \quad \forall j \]

By extension, we can identify the second best matching unit, the third, and so on. We apply a standard weight update rule with a learning rate that decays over time, \(\alpha(t) = 0.05/(1 + t/2000)\) and a Gaussian neighbourhood function of the distance between neurones \(i\) and \(k\), that shrinks in time \(N(i, k) = e^{-||r_i - r_k||^2/2\sigma^2(t)}\). We define an averaged quantisation error, as a measure of weight alignment to the input, so that the Euclidian distance between input patterns and their respective best matching unit is:

\[ E = \langle ||x - m_c(x)|| \rangle_x \]

where \(m_c(x)\) is the best matching unit for input pattern \(x\). In order to shorten simulation time in experiments 1 and 3, we used a batch version of the algorithm.\(^8\) In all experiments, map sizes were fixed to a 9x12 hexagonal grid of neurones for the visual map and to a 5x7 grid for the acoustic map.
2.2. Coding the inputs

2.2.1. Image generation

Images in the dataset are created from six pictures of animals (dog, cat, cow, fish, pig, sheep). Each image is first bitmapped into a square image having 20x20 pixels. We generate blurred versions of the six pictures in order to create multiple tokens in each category, centred on a prototype. We create 18 images per category by changing the grey scale value of a random number of pixels (min 0: max 400). The magnitude of the grey scale change is drawn from a normal distribution centred on zero and a standard deviation equal to 80% of the full grey scale. Prototypes are not included in the data set.

2.2.2. On the importance of real acoustic token

There is little consensus in the field as to what acoustic information babies use when identifying words. A series of studies emphasise the fact that babies pay attention to much more than simple features that would be described by a simple phonological encoding. In particular, it has been shown that at 9 months of age, babies are sensitive both to stress and phonetic information,\(^{12}\) at 9.5 months they are able to make allophonnic distinctions\(^{13}\) and at 17 months, they pay attention to co-articulation.\(^{14}\) All of these sensitivities to the speech signal may have an important impact on early lexical development. Therefore, we exploit the whole acoustic signature of tokens in order to avoid discarding relevant acoustic information.

We extract the acoustic signature from raw speech waveforms for six acoustic categories produced by nine female native speakers. By doing so, we confront the model with a lack of invariance in word pronunciation introduced by different speakers. Tokens are then normalised in length and sampled at regular intervals, 3 times per syllable\(^{a}\). After sampling, the sounds are filtered using the Mel Scale in order to approximate the human ear sensitivity. Input vectors are concatenations of three 7-dimensional mel-cepstrum vectors, derived from FFT-based log spectra\(^{b}\).

\(^a\)We found that for monosyllabic words, having 2 samples per syllable is sufficient from the point of view of word-object generalisation performance as described in the results section. We found no statistically significant improvement when increasing the number of time-slices beyond \(N = 2\).

\(^b\)The mel-cepstrum vectors are obtained by applying the following procedure: take the Fourier transform of a windowed excerpt of the signal, map the log amplitudes of the spectrum obtained above onto the Mel scale, using triangular overlapping windows and
2.3. Forming the cross-modal associations

After maps are structured following the presentations of the images and acoustic tokens in the data set, we mimic joint attentional activities between the care-giver and the baby by presenting simultaneously to both maps a randomly picked image from the data set and an acoustic token randomly picked within the matching category (e.g. one of the 18 images of dogs and one acoustic signature of a speaker saying the word “dog”). We build cross-modal associations by learning Hebbian connections between both maps. As a further simplification of the model, we use bidirectional synapses whose amplitudes are modulated by the activity of the connecting neurones. We define the neural activity of a neurone $k$ to be $a_k = e^{-q_k/\tau}$ where $q_k$ is the quantisation error associated with neurone $k$ and $\tau = 5$ is a normalisation constant. There are several options for linking the maps:

- link all neurones on both maps
- link only the Best Matching Unit of the paired image on the visual map and the Best Matching Unit of the paired acoustic token on the acoustic map
- link together only a percentage of the neurones on both maps.

In experiments 1 and 2, only the top 25% of the Best Matching Units are linked together whereas in experiment 3 the percentage of units that are allowed to fire and wire is varied in order to investigate the role of this linking parameter when generalising word-object associations.

All synapses were first randomly initialised with a normal distribution centred on 1 and with a standard deviation of $\frac{1}{\sqrt{1000}}$. Synapse amplitudes are modulated according to a standard Hebb rule with saturation. Therefore synapse weights stay in a physiological range even for high neural activities. The synapse connecting neurone $i$ from the visual map to the neurone $j$ of the acoustic map is computed as follows:

$$w_{ij}(n+1) = w_{ij}(n) + 1 - e^{\lambda a_i a_j}$$

where $n$ refers to the index of the word-object pairing and $\lambda = 10$ is the learning rate. The free parameters $\tau$ and $\lambda$ were chosen by inspection to provide good results. After every word-object presentation, weights are normalised so as to model the limited synaptic resources: $\sum_{ij} w_{ij}^2 = 1$.

After training on cross-modal pairings we assess the capacity of the network to extend the association of a presented word-object pair to non-paired

finally take the Discrete Cosine Transform of the list of Mel log-amplitudes.
items that belong to the same category. Following a number of simultaneous presentations of word-object pairs, weights are fixed and all images in the dataset are classified according to whether the induced activity on the acoustic map corresponds to the activation of the appropriate label. This is referred to as the visual to acoustic condition: \(v2a\). Averaging over all images in the data set gives us a measure which we call the classification success, \(C\). We compare this measure to:

- the perfect classification \(C_{\text{max}}\), achievable given the number of pairings (perfect classification in categories that “posses” a pairing, random classification in other categories)
- item-based classification \(C_{\text{item}}\), where the items presented in pairs are associated perfectly, the other ones being classified at random
- the baseline where no learning occurs, the random guess, with one chance out of six to classify correctly the image.

We define a normalised value for generalisation, \(G\), so that perfect classification given the pairings has a value of 1 and perfect memory with no generalisation, the item-based condition, would give a score of 0:

\[
G = \frac{C - C_{\text{item}}}{C_{\text{max}} - C_{\text{item}}}
\]

Similarly acoustic tokens are classified according to the activity induced onto the visual map, referred to as the acoustic to visual condition: \(a2v\). All results reported are averaged over 65 independent simulations.

3. Results

3.1. Generalisation as a function of number of pairings

We report both classification success and its normalised version, the generalisation measure, as a function of the number of word-object pairs on which the network has been trained. We expect the network classification success to improve with an increasing number of label-object pairs. A positive correlation between classification success and the number of joint presentations of objects and their labels is shown in the left panel of Fig. 1. In both conditions, classifying images by their induced activity onto the acoustic map (\(v2a\) condition) and classifying labels by monitoring the induced neural activity onto the visual map (\(a2v\) condition), the network outperforms their respective “no generalisation” baselines. This indicates that even following the presentation of a single word-object pair, the network is capable of
generalising the association to other objects and labels within the same category. The difference between conditions in overall classification success can be explained by the different levels of variance associated with the visual and the acoustic inputs.

Fig. 1. Correct associations of labels to objects as a function of the number of simultaneous presentations of word-object pairs, after maps are structured. Left panel: Classification success for both conditions results are compared both to the maximal classification achievable (solid line) and to the results of a system only learning associations between paired tokens, with no generalisation capacity (dashed line for the a2v condition and dash-dotted line for the v2a condition). The dotted line corresponds to a random association of word and objects. Right panel: Generalisation as a function of the number of word-object pairings. Error bars correspond to one standard deviation after averaging over 65 simulations.

The right-hand panel of Fig. 1 depicts the normalised counterpart of classification success. We notice that to a first approximation there is no strong dependence upon the number of training pairs. In other words, when compared to both an optimal generalisation and to a simple item-based learning device, the network has a constant generalisation capacity. Even though the absolute ability of the network to associate objects with their labels increases along with the amount of joint word-object experience, the relative capacity of the network to generalise such associations is essentially independent of the number of such pairings. This indicates that the capacity of the network to generalise is dependent on the neural architecture, both at the level of the quality of the organisation of the maps and on the number
of Hebbian associations that are allowed to fire and wire together. In the next experiment, we investigate the role played by the map qualities for the network’s generalisation ability.

### 3.2. Generalisation as a function of pre-pairing experience

In order to investigate the role played by the neural architecture for generalisation capacity, we first controlled the quality of the maps structure before presenting the network with word-object pairs. Map structure improves with the experience. We monitored the average quantisation errors of both maps as a function of the number of times the whole data set is presented to the maps (defined as an epoch). In the bottom left panel of Fig. 2 we see the monotonic decline in quantisation errors for both maps as a function of increasing experience of images and sounds. In the top left panel of Fig. 2 we plot the classification success for the a2v condition as a function of pre-pairing experience, after training on 12 word-object pairs. The top right-hand panel plots the same measure for the v2a condition.

In both panels, the classification success curves start from a random guess baseline to cross a level of performance comparable to that of an item-based learning network, before reaching a plateau in classification performance. When maps are still unstructured, neural activity is very low when an item is presented (the quantisation error is high). Hence, Hebbian learning is still too weak to be able to associate paired items reliably. When structure starts to emerge, presentation of word-object pairs elicit map activities sufficiently large to promote significant weight changes. At this stage in map development, object-label pairs are associated item-by-item but the lack of topological organisation in the maps is such that generalisation cannot be sustained yet. This stage of item-based performance corresponds to 100 epochs in the a2v condition and 35 epochs in the v2a condition. Finally, when the network has enough experience with items before pairings, maps are well organised and associations made during the pairing phase are generalised well to the other non-paired items.

In the bottom right panel of Fig. 2 we plot classification success as a function of the simple average of the quantisation errors of the maps. This comparison provides a more direct index of the impact of map structure on generalisation. The monotonic increase in generalisation quality as the maps structural quality increases (quantisation errors decrease) confirms

---

*Very similar results were obtained with 4 word-object pairs.*
Fig. 2. Top row: classification success as a function of pre-pairing experience with objects and labels in the a2v condition (left) and v2a condition (right). Bottom left panel: quantisation error (map structure) as a function of experience. Bottom right panel: classification success as a function of quantisation error (map structure).

our claim that generalisation quality ultimately depends on the pre-lexical (pre-pairing) categorisation abilities.

3.3. Generalisation capacity as a function of the number of Hebbian associations

Finally we investigate the role played by the number of neurones allowed to be associated through Hebbian connections. Fig. 3 displays classification success as a function of the percentage of the neurones that are allowed to fire and wire together. When only one Best Matching Unit on each map is allowed to fire and wire, generalisation of labels to other images (and of images to other labels) fails. By increasing the number of nodes allowed to connect on both maps, we reach a maximal generalisation capacity. It is noteworthy that performance reaches a plateau when about 15 to 25% of the neurones are connected to each other. Additional capacity does not result in improved generalisation. The slow decay in the quality of generalisation past this maximum is explained by the penalty induced by introducing a
greater number of weights.

![Classification success as a function of the percentage of maps that are linked through Hebbian connections.](image)

Fig. 3. Classification success as a function of the percentage of maps that are linked through Hebbian connections.

It might be argued that an autonomous procedure designed to identify the Best Matching Units is not biologically plausible. However, there are several potential solutions to this problem. First, the synapses that need to be reinforced are precisely those connecting neurones simultaneously with high activities. It is not unreasonable to suppose that a natural pruning procedure would eliminate the silent synapses so that only the strong ones would survive. This way only a limited fraction of nodes would be connected through associative links. Alternatively, because of the topological organisation of the SOMs, neurones representing similar items are close together. Hence, the BMUs are in the same region of the map and would not require a complex search procedure, satisfying both a limited synaptic resource argument and a locality rule.

4. Discussion

Connectionism is often considered to be an implementation of exemplar-based learning procedures. The way that supervised networks generalise novel inputs is through interpolation from the training set. Therefore, it is not guaranteed that such models will obey the taxonomic constraint. However, in our model, the first phase of training is completely unsupervised and we show that a single label-object pair can yield taxonomic responding
The mechanism driving taxonomic responding is closely related to the percentage of neurones that are allowed to fire and wire (Exp. 3). If only one BMU from each map is allowed to fire and wire, taxonomic responding is not achieved. However, only a limited amount of synaptic resources is required in order to support good generalisation; we demonstrated in Exp. 3 that only about 20% of the maps need to fire and wire to achieve taxonomic responding.

The other prerequisite for good word-object generalisation in our model is to provide the system with well structured maps. In other words, successful word-object associations and their generalisations rely on pre-existing categorisation abilities. Once the perceptual system can achieve coherent object and sound categorisation, word-object associations are learnt fast and generalise well.

This finding is consistent with a series of studies suggesting that speech perception and cognitive development in infancy predicts language development in the second year of life.\textsuperscript{16–19} Similarly, deficits in speech perception (bad auditory categorisation) predict language learning impairments\textsuperscript{20} and more generally both delayed auditory perception\textsuperscript{21} and impairments of visual imagery\textsuperscript{22} are correlated with specific language impairments. The model captures the claim made that auditory perception bootstraps word learning.\textsuperscript{23} We might also point out that the model predicts that visual categorisation bootstraps word learning in a similar fashion. Our findings suggests that generalisation of object-label associations are dependent upon good pre-lexical categorisation abilities and offers theoretical support for the experimental findings\textsuperscript{16,17} that improvements in speech perception during infancy is an important developmental step toward language acquisition.

In summary, we show from a modelling perspective how taxonomic responding can be built on pre-existing categorisation abilities, along with limited synaptic resources. This neuro-computational account of taxonomic responding confirms the importance of pre-lexical categorisation abilities as predictors of successful lexical development.

References