

# Multi-sensory information as an improvement for communication systems efficiency

Lacerda F., Klintfors E., Gustavsson L.

Department of Linguistics, Stockholm University

## Abstract

*The paper addresses the issue of extraction of implicit information conveyed by systematic audio-visual contingencies. A group of adult subjects was tested on a simple inference task provided by short film sequences. The video materials were encoded and submitted to processing by two neural networks (NN) that simulated the results of the adult subjects. The results indicated that the adult subjects were extremely efficient at picking up the underlying information structure and that the NN could also perform acceptably on both classification and generalization tasks.*

## Introduction

Language acquisition can be described as a process through which infants derive the underlying linguistic structure of their ambient language. In spite of the complexity and variability of the language input it is an undeniable fact that within about two years of life, typical infants are able to pick up the linguistic regularities of the ambient language. Making sense of linguistic information that is implicitly conveyed in a diversity of speech communication situations appears to be such an insurmountable task that researchers are prone to consider that some sort of initial guidance is necessary to home in on the ambient language's underlying principles (Chomsky, 1968; Pinker, 1994). The present paper attempts to challenge this established view by sketching a scenario in which linguistic information may be derived in the absence of pre-knowledge or dedicated linguistic biases. Indeed language can be seen as an emergent consequence of the interplay between the infant and its environment, where the richness and structure of the sensory flow may contain enough information to trigger language development (Jusczyk, 1985; Elman, Bates, Karmiloff-Smith, Parisi, & Plunkett, 1997). More explicitly the language acquisition hypothesis to be tested in this paper relies on the assumption that linguistic structure is implicit in the

multi-sensory information available to the young infant (Lacerda et al., 2004a; Lacerda, 2003; Lacerda & Lindblom, 1997). In the early stages of language acquisition the infant and the adult tend to communicate about objects or occurrences in the infant's immediate neighbourhood. Although the speech signal that the infant is exposed to may indeed refer to absent objects or abstract concepts, the gist of the infant-directed speech tends to be focused on very tangible objects that the adult assumes to be in the infant's focus of attention. Under such ecologically relevant scenarios of adult-infant interaction, there is an almost inevitable correlation between what the infant hears and its visual, tactile or gustative sensations. In other words, because spoken language is used to refer to objects or events, the sound structure of the speech signal representing those referents must be highly correlated to at least some of the sensory representations of the objects or events it refers to. As a consequence, the very co-occurrence of certain sequences of speech sounds and sensory representations of the objects they are associated with conveys implicit information on the speech signal's linguistic referential function. To be sure, the relationship between the auditory representation of the speech signal and the sensory representations of its referents is far from being deterministic. There is no guarantee that a given instance of speech signal will be referring to the object that happens to be in the young infant's field of attention. On the other hand, because the assumption of poverty of stimulus implies that probability of recurrent matches between the auditory representation of the speech signal and the sensory representation of its referent is vanishingly small given language's potentially unlimited combinatorial possibilities, even a barely resembling repetition of the co-occurrence of the speech signal with its referent is extremely significant. Indeed repetition of shorter or longer utterances is the hall mark of speech directed to very young infants (Lacerda et al., 2004b).

### Combining Auditory and Visual Information in a neural network (NN) model

The NN model in this study combines visual and auditory information. The model is based on data used to test adults' spontaneous propensity to extract latent information from a short video sequence. Tests on young infant's ability to extract referential information from audio-visual contingencies are also being conducted and will be reported later.

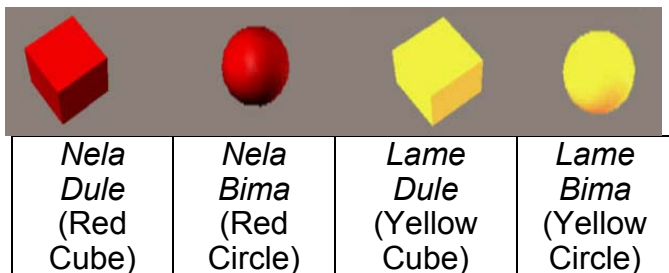


Figure 1. Illustration of the film sequence shown to the adults. The four pictures demonstrate 6 sec long film sequences each. The objects moved smoothly across the screen while sound track played two repetitions of the two-word sentences formed by the non-words that had been arbitrarily assigned to represent the colors and the shapes of the objects.

The videos shown to the adults (figure 1) were about 24 seconds long and consisted of four sequences: first a red cube was shown moving smoothly across the screen, then a red circle performed the same motion, the third sequence showed a yellow circle also following the standard path. Two-word nonsense sentences, created by concatenating nonsense words that were arbitrarily assigned to represent the color and the shape of the objects, were played along these figures. After exposure to the materials the adult subjects were presented with answer sheet were questions concerning the meaning of the nonsense words used in the videos were embedded in a number of foils containing spurious words and situations.

#### Reference data from the adult subjects

A group of 21 adult subjects participated in the simulated "language learning" experiment described above. The subjects were asked to sit in front of a blank computer screen and without further instructions the video sequences were started. After this exposure the subjects were asked to describe what they just had seen and heard. Most of the subjects referred spontane-

ously to the embedded meanings of the non-words. After this first phase the video sequence that the subject just had seen was played in loop at the same time that the subject attempted to answer multiple choice questions concerning the nonsense names of the colors and the shapes of the objects. Whatever responses produced by the subjects were used as indications of the learned sound-object attribute pairing. The task was obviously very simple for the adult subjects and it turned out that the vast majority of the responses simply reflected the built-in contingencies embedded in the stimuli. The results of 21 subjects showed 97.6% correct generalization (see figure 2). The errors made by two subjects were incorrectly named shapes of the objects. The colour attribute generalization was 100% correct.

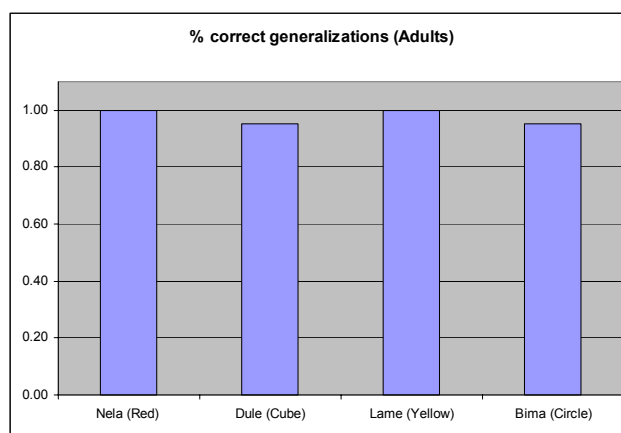


Figure 2. Reference data provided by the adult subjects. Percentage correct discoveries of the meaning of the non-words representing the colours and the shapes of the objects. Only in two cases were there errors made by the subjects.

#### The auto association model

To carry out NN simulation, two types of feed-forward architectures were constructed. The first – an auto association model (figure 3) – was intended to simulate *a priori* knowledge, the task of the NN simply being to associate the colours and the shapes of the visual objects to the non-words corresponding to these two attributes, i.e. to reproduce its input at the output level. Exactly the same set of data formed the input and the output patterns of the NN. The 96-bit input vectors encoded the visually shown colour (bits 1 to 24), the visually shown shape (bits 25 to 48), the auditorily presented non-word standing for the colour (bits 49 to 72), and the auditorily presented non-word standing

for the shape (bits 73 to 96). The hidden node layer consisted of two hidden units.

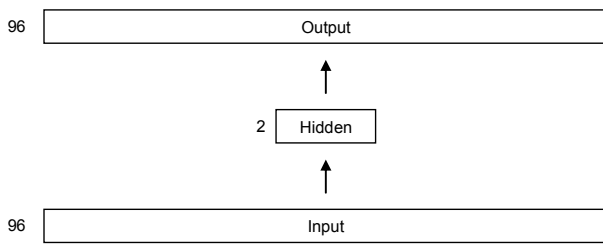


Figure 3. Schematic architecture of the auto association NN aimed at simulating a priori knowledge. The task of the NN was to reproduce its input at the output level. The number of units in each layer is indicated by the figures to the left of rectangles.

In the first run of the NN, the information from the input layer was separated in two different receptive fields – one of them corresponding to vision and the other to audition. In the second run of the NN all the information from the input layer was passed to each of the two hidden nodes. Training was done sequentially by presenting the model with a visual colour and shape parameter set and its associated non-word auditory set. The NN was able to find redundancies in the distributed data and accordingly all the input patterns were correctly categorized. The performance of the NN was stable regardless of the way of passing input information to the hidden units.

### The generalization model

The second architecture of the NN (figure 4) attempted to simulate the fact that the adult subjects in addition to being able to discriminate the colours and the shapes of the objects, also learned the concepts conveyed by the non-words and were readily able to apply them to new contexts. This NN also had 96 bit-vectors as input. The hidden node layer consisted of four hidden-nodes, the first and the second of them receiving information from the visual half of the input nodes, and the third and the fourth of them receiving information from the auditive half of the input nodes. The NN was in this way structured in two input channels so to simulate two kinds of correlated sensory input.

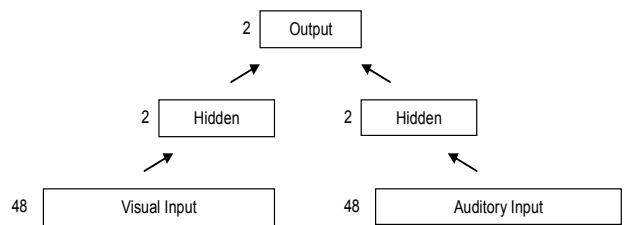


Figure 4. Schematic architecture of the NN with two sensory input channels aimed at simulating learning. The task of the NN was to generalize its knowledge of colours/shapes of figures and show it via recognizing familiar colours despite of novel figures, as well as recognizing novel-coloured familiar figures.

The performance of the NN was tested with help of data not previously shown to the NN. The novel data consisted of a blue cube and a green circle (new colours for familiar shapes) as well as of a red cone and a yellow cylinder (new object shapes for familiar colours), as illustrated in figure 5.

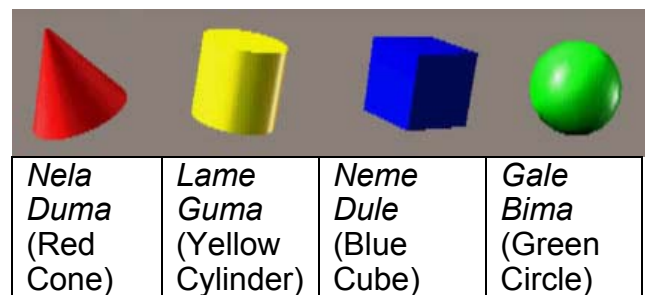


Figure 5. Illustration of the novel data used to test the NN's ability to generalize its knowledge in new context. Before the experiment these non-words were arbitrarily assigned to represent the novel colours and the novel shapes of the objects.

The output activations showed that the NN was able to – not only associate the non-words corresponding to visually presented colours and shapes of the objects – but also to generalize its knowledge to a new context. In a second run of the NN all the input information was passed to each of the four units at the hidden-layer. Just as in the above auto association simulation, the results of this run were not affected by the change of the way of passing information to the hidden units.

## Discussion

The ultimate goal of this study is to investigate how human infants might be able to extract im-

PLICIT information from their experience with audio-visual stimuli. At this stage we simply ran a pilot study using adult subjects that were asked to watch short video sequences and subsequently requested to answer questions related to the implicit information potentially conveyed by the audio-visual stimuli. Although the adult subjects did not receive any instructions (in an attempt to make the situation more comparable to that of the infants') the subjects had no difficulties in inferring the underlying structure right on the first inquire. The situation created by these stimuli was obviously too simple for the adult subjects, who could not avoid thinking of the goals and the structure of the stimuli as soon as they were put in the experimental situation. Our next question concerns the extent to which the infant subjects may also be able to detect and generalize the implicit audio-visual consistencies. Although we still do not have data from the infants and we expect the infants' performance to be age-dependent, it is reassuring that NN's performance mimic so well the adults' results. This leads us to envisage the future infant speech perception experiments as a means to evaluate the potential significance of NN models in accounting for the grounds of linguistic development departing from general-purpose association mechanisms. In general the outcome of a NN is dependent on its architecture but our results do not suggest critical dependence on any of the two architectures tested.

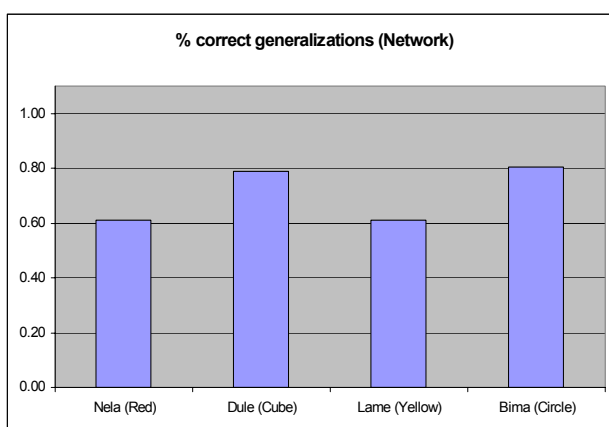


Figure 6. Results of the NN performance. Percentage correct generalizations of the meaning of the non-words representing the colours and the shapes of the objects. The results indicate that generalization of shapes was slightly more robust than generalization of colours.

To be sure, the stimuli used in this first experiment are likely to be too simple to fully demonstrate relevant language development relying on general-purpose associative mechanisms. Therefore our current experiments with infants are being conducted using audio-visual contingencies that attempt to replicate ecologically relevant communication settings.

## Acknowledgements

This work was supported by grants from the Swedish Research Council, the Bank of Sweden Tercentenary Foundation and Birgit & Gad Rausing's Foundation.

## References

- Chomsky N. (1968). *Language and mind*. New York: Harcourt Brace Jovanovich.
- Elman J., Bates E., Karmiloff-Smith A., Parisi D., & Plunkett K. (1997) *Rethinking innateness*. Cambridge, Massachusetts: MIT Press.
- Jusczyk P. (1985) On characterizing the development of speech perception. In Mehler J. & Fox R. (eds), *Neonate cognition: Beyond the blooming, buzzing confusion* Hillsdale, New Jersey: Lawrence Erlbaum, 199–299.
- Lacerda F. (2003) Phonology: An emergent consequence of memory constraints and sensory input. *Reading and Writing: An Interdisciplinary Journal*, 16, 41–59.
- Lacerda F., Klintfors E., Gustavsson L., Lagerkvist L., Marklund E., & Sundberg U. (2004a) *Ecological Theory of Language Acquisition*. In Genova: Epirob 2004, 147–148.
- Lacerda F. and Lindblom B. (1997) *Modelling the early stages of language acquisition*. In Olofsson Å. and Strömquist S. (eds), *Cross-linguistic studies of dyslexia and early language development*. Office for official publications of the European Communities, 14–33.
- Lacerda F., Marklund E., Lagerkvist L., Gustavsson L., Klintfors E., & Sundberg U. (2004b) *On the linguistic implications of context-bound adult-infant interactions*. In Genova: Epirob 204, 149–150.
- Pinker S. (1994) *The Language Instinct: How the Mind Creates Language*. (1 ed.) New York: William Morrow and Company, Inc.