An Ecological Theory of Language Acquisition (*)

Francisco Lacerda
frasse@ling.su.se
Department of Linguistics, Stockholm University (Sweden)

Ulla Sundberg
ulla@ling.su.se
Department of Linguistics, Stockholm University (Sweden)

ABSTRACT. An ecological approach to early language acquisition is presented in this article. The general view is that the ability of language communication must have arisen as an evolutionary adaptation to the representational needs of Homo sapiens and that about the same process is observed in language acquisition, although under different ecological settings. It is argued that the basic principles of human language communication are observed even in non-human species and that it is possible to account for the emergence of an initial linguistic referential function on the basis of general-purpose perceptual, production and memory mechanisms, if there language learner interacts with the ecological environment. A simple computational model of how early language learning may be initiated in today's human infants is proposed.

KEY-WORDS. Infancy; Early language acquisition; Babbling; Infant-Directed Speech.

(*) Acknowledgements. This work was supported by project grants from The Swedish Research Council, The Bank of Sweden Tercentenary Foundation (MILLE, K2003-0867) and EU NEST program (CONTACT, 5010). The first author was also supported by a grant from the Faculty of Humanities, Stockholm University, and the second author received additional support from The Anna Ahlström and Ellen Terserus Foundation. The authors are also indebted to the senior members of the staff at the Dept. of Linguistics, Stockholm University, for their critical and helpful comments on earlier versions of this manuscript.

Introduction

The ability for language communication is a unique human trait differentiating us from all other species, including those most closely related in genetic terms. In spite of the morphological and molecular proximity to closely related species in the family Hominidae, like *Pongo* (orangutan), *Gorilla* (gorilla) and *Pan* (chimpanzees) (Futuyama 1998: 730), it is obvious that only *Homo* (human), and most likely only *Homo sapiens*, has evolved the faculty of language. In evolutionary terms, the capacity for language communication appears to be a relatively recent invention that nowadays is "rediscovered" by typical human infants. From this broad perspective it may be valuable to investigate parallels between the phylogenetic and ontogenetic components of the process of language communication, while keeping in mind that Ernst Haeckel's (1834-1919) biogenetic law, "ontogeny recapitulates phylogeny" (1866), is indeed far too simplistic to be taken in strict sense (Gould 1977). This article opens with a broad look at the evolutionary aspects presumably associated with the discovery of language as a communication system. This evolutionary perspective will focus on general notions regarding the emergence of the language communication ability in the *Homo*'s ecological system. In the reminder of the article, an ecologically inspired developmental perspective on early language acquisition in human infants will be presented, enhancing potential similarities and differences between the developmental and evolutionary accounts.

From the phylogenetic perspective, language communication, in the sense that it has today, may have emerged about 200 to 300 thousand years ago (Gärdenfors 2000) with the advent of *Homo sapiens*[1]. Until then it is assumed that hominids might have communicated mainly by calls and gestures. Derek Bickerton (Bickerton 1990), for instance, speculates that a protolanguage, assigning referential meaning to arbitrary vocalizations, may have been used already by *Homo erectus*, about 1.5 Mya to 0.5 Mya (Futuyama 1998: 731). Such a protolanguage must have been essentially a "referential lexicon"

[1] "Most hominid fossils from about 0.3 Mya onward, as well as some African fossils about 0.4 My old, are referred to as *Homo sapiens*." (Futuyama 1998).

and somewhat of a "phylogenetic precursor of true language that is recapitulated in the child (…), and can be elicited by training from the chimpanzee" (Knight, Studdert-Kennedy & Hurford 2000: 4). While this protolanguage may have consisted of both vocalizations and gestures used referentially, the anatomy of the jaw and skull in earlier hominids must necessarily have constrained the range of differentiated vocalizations that might have been used referentially. Considerations on the relative positions of the skull, jaw and vertebral column in early hominids clearly indicate significant changes in the mobility of those structures, from the *Australopithecus afarensis* (4.0-2.5 Mya) to *Homo sapiens* (0.3 Mya), changes that must have had important consequences for the capacity to produce differentiated vocalizations. An important aspect of these changes is the lowering of the pharynx throughout the evolutionary process. This lowering of the pharynx is also observed on the ontogenetic time scale. Newborn human infants have at birth a very short pharynx, with high placement of their vocal folds, but they undergo a rather dramatic lengthening of the pharynx within the first years of life, in particular within about the first 6 months of life. As will be discussed later in this article, this short pharynx in relation to the oral cavity has some important phonetic consequences that can be summarized as a reduction in the control that the infant may have over the acoustic properties of its vocalizations. And from the acoustic-articulatory point of view, there are essentially no differences between the expected acoustic-articulatory relationships in the infant or in the adult since it seems that the infant's vocal tract in this case can be considered as a downscaled version of an *Australopithecus afarensis* adult. A more difficult question is to determine at which point of the evolutionary process did the laryngeal position reach the placement it has nowadays but the general consensus is that *Homo erectus* probably marks the transition from high to low laryngeal positions (Deacon 1997: 56). However, in evolutionary terms the emergence of a new anatomical trait does not directly lead to the long-term consequences that are found retrospectively. Changes are typically much more local, opportunistic and less dramatic than what they appear to be on the back-mirror perspective, reflecting Nature's tinkering rather than design or goal-oriented strategies (Jacob 1982) and *Homo erectus* probably did not engage in speech-like communication

just to explore the new emerging anatomical possibility. A more plausible view is then that *Homo erectus* might have expanded the number of vocalizations used to signal objects and events available in the ecological environment, nevertheless without exploring the potential combinatorial and recursive power of the vocalizations' and gestures' inventories. Indeed, as recently pointed out by Nishimura and colleagues (Nishimura, Mikami, Suzuki & Matsuzawa 2006), the descent of the larynx is observed also in chimpanzees, leading to changes in the proportions of the vocal tract during infancy, just like in humans, and the authors argue that descent of the larynx may have evolved in common ancestor and for reasons that did not have to do with the ability to produce differentiated vocalizations that might be advantageously used in speech-like communication. In particular with regard to gestures, it is also plausible that *Homo erectus* might not have used gestures in the typically integrated and recursive way in which they spontaneously develop in nowadays *Homo sapiens*' sign language (Schick, Marschark, Spencer & Ebrary 2006), although their gestural anatomic and physiologic capacity for sign language must have been about the same. Thus, in this line of reasoning the question of how nowadays language communication ability did arise is not settled by the lowering of the larynx, per se, and the added differential acoustic capacity that it confers to vocalizations.

Since communication using sounds or gestures does not leave material fossil evidence it is necessary to speculate how vocalizations might have evolved into today's spoken language and draw on useful parallels based on the general communicative behaviour among related hominid species. Situation-specific calls used by primates, such as the differentiated alarm calls observed in Old World monkeys like the vervet monkeys (*Cercopithecus aethiops*) (Cheney & Seyfarth 1990), may in fact be seen as an early precursor of humanoid communication. To be sure, the calls used by the vervet monkeys to signal different kinds of predators and prompt the group to take appropriate action are rather primitive as compared to the symbolic functions of modern human language communication but they surely demonstrate a multi-sensory associative process from which further linguistic referential function may emerge. The ability to learn sound-meaning relationships is present in primates at a variety of levels and

animals tend to learn the meaning of different alarm calls, even those used by other species, if those calls provide them with meaningful information and they can even modify their vocalizations to introduce novel components in the established sound-meaning coding system (Gil-da-Costa, Palleroni, Hauser, Touchton & Kelley 2003; Hauser 1989, 1992; Hauser & Wrangham 1987). But using multi-sensory associations does not pay-off in the long run under the pressure of increasingly large representational demands and once the need for representation of objects or events in the ecological setting reach a critical mass of about 50 items, sorting out those items in terms of rules becomes a much more efficient strategy (Nowak, Plotkin & Jansen 2000; Nowak, Komarova & Niyogi 2001). In line with this, the driving force towards linguistic representation and the emergence of syntactic structure appears to come from increased complexity in the representation of events in the species ecological environment and again there are plausible evolutionary accounts for such an increase in representational complexity with the advent of *Homo erectus*. One line of speculation is linked to bipedalism. Bipedal locomotion is definitely not exclusive of humans (avian species have also discovered it and use it efficiently, like in the case of the ostrich) but provided humanoids with the rather unique capacity of using the arms for foraging and effectively carrying food over larger distances. Again, mobility, per se, does not account for the need to increase representational power (migration birds travel over large distances and must solve navigation problems and yet have not developed the faculty of language because of that) but it adds to the complexity of the species ecological setting, posing increasing demands on representational capacity. Another type of navigational needs that has been suggested as potentially demanding more sophisticated representation and communication abilities is the more abstract navigation in social groups (Tomasello & Carpenter 2005) and the abstract demands raised by the establishment of "social contracts", as Terrence Deacon had pointed out in his 1997 book (Deacon 1997). To be sure, abstract representation capacity can even be observed, at least to a certain extent, in non-human primates (Savage-Rumbaugh, Murphy, Sevcik, Brakke, Williams & Rumbaugh 1993; Tomasello, Savage-Rumbaugh & Kruger 1993) but nevertheless they fall short of human language as far as syntax and combinatorial ability is concerned.

The general notion conveyed by this overview of the possible evolution of language communication is that the capacity for language communication must have emerged as a gradual adaptation, building on available anatomic and physiologic structures whose functionality enables small representational advantages in the humanoids' increasingly complex ecological contexts. Of course, because evolution is not goal-oriented, the successive adaptations must be seen more as accidental and local consequences of a general "arms-race" process that although not aimed at language communication still got there, as it can be observed in retrospective. Evolutionary processes are typically like this, lacking purpose or road-map but conveying a strong feeling of purposeful design when looked upon in the back-mirror. William Paley's proposal of the "watchmaker argument" is an example of how deep rooted the notion of intelligent design can be when viewing evolutionary history in retrospective (and knowing its endpoint from the very beginning of the reasoning). Yet it is well known that complex or effective results do not have to be necessarily achieved by dramatic or complex design, but that they are rather often the consequence of deceivingly simple interactions that eventually produce relatively stable long-term consequences (Enquist & Ghirlanda 2005; Dawkins 1987, 1998). The problem is accepting non-teleological explanations when the final results appear to be so overwhelmingly clear in their message of an underlying essential order. But how could that be otherwise? How could the current state of the evolutionary process not look like a perfectly designed adaptation to the present ecological settings if non-adapted behaviour or structures confer disadvantages that undermine the species' survival? Obviously the potential advantages that small and purposeless anatomic and physiological changes might confer are strictly linked to the ecological context in which they occur. Suppose, for instance, that our recently discovered ancestor, *Tiktaalik* (Daeschler, Shubin & Jenkins 2006), the missing link between fish and tetrapods, had appeared now rather than under the Devonian-Carboniferous Period, between 409 Mya and 354 Mya (Futuyama 1998): How long would Tiktaalik have survived in today's ecological settings? Probably not very long. The bony fins that the animal successfully used as rudimentary legs conveyed significant advantage in an ecological context where land life was limited to

the immediate proximity of water and therefore there were no land competitors. Under these circumstances, the ability to move on the water bank just long enough to search for fish that might have been trapped in shallow waters was obviously a big advantage but would have made the animal an easy prey nowadays, unless it had evolved shielding, poisoning or stealth strategies to compensate for its low mobility. In other words, the potential success of traits emerging at some point in the evolutionary process is intimately dependent on the ecological context in which they appear, suggesting that even a rudimentary ability to convey information via vocalizations may have offered small but systematic and significant advantages to certain groups of hominids. Indeed, an important aspect of this ecological context is that the innovation presented by even a rudimentary discovery of representational principles is likely to spread through cultural evolution processes (Enquist & Ghirlanda 1998; Enquist, Arak, Ghirlanda & Wachtmeister 2002; Kamo, Ghirlanda & Enquist 2002), propagating the innovation to individuals that may be both upstream, downstream or peers in the species' genetic lineage. From this perspective language evolution must be seen as the combined result of both cultural and genetic evolutionary processes.

The ability to use vocal or sign symbols to represent events or objects in the ecological context was probably common at least within several hominid species (Johansson 2005) but its recursive use must have been discovered by *Homo sapiens* whose brain capacity developed (Jones, Martin & Pilbeam 1992) in an "arms-race" with the increasingly complex symbolic representational demands. From this evolutionary perspective, the language acquisition process observed in nowadays infants can be seen as a process that in a sense repeats the species' evolutionary story, keeping in mind that it starts off from an ecological context where language communication per se does not have to be re-discovered and is profusely used by humans in the infant's immediate environment. Of course, since both the early *Homo sapiens* and today's infants adjusted and adjust to the existing ecological context, the dramatic contrast of the communicative settings of their respective ecological scenarios must account for much of the observed differences in outcome. The challenge here is to account for how this language acquisition process may unfold as a consequence of the general interaction between the infant and its ambient.

The infant's pre-conditions for speech communication

To appreciate the biological setting for language acquisition it is appropriate to take a closer look at some of the infant's capabilities at the onset of life.

Among the range of abilities displayed by the young infant, perceiving and producing sounds are traditionally considered to be the most relevant to the language acquisition process. These capabilities are not the only determinants of the language acquisition process (language acquisition demands multi-sensory context information, as it will be argued below) but they certainly play an important role in the shaping of the individual developmental path. If there are biologically determined general propensities to perceive or generate sounds in particular ways, these propensities are likely to contribute with significant developmental biases that ought to become apparent components of language acquisition, although such biological biases are dynamic plastic components of the language acquisition process that will necessarily influence and be influenced by the process itself (Sundberg 1998) but here the focus will be on just some production and perception biases.

Production constraints

To estimate the infant's production strategies, one may use the evidence provided by infant vocalizations in order to derive underlying articulatory gestures. However the acoustic analysis of high pitch vocalizations is not a trivial task, since high fundamental frequencies effectively reduce the resolution of the available spectral representations. An additional difficulty is caused by the anatomic differences between the infant's and the adult's vocal tracts. The vocal tract of the newborn infant is not a downscaled version of the adult vocal tract. One of the most conspicuous departures from proportionality with the adult vocal tract is the exceedingly short pharyngeal tract in relation to the oral cavity observed in the newborn infant (Fort & Manfredi 1998). If the infant's vocal tract anatomy was proportionally the same as the adult, the expected spectral characteristics of the infant's utterances would be essentially the same as the adult's although proportionally shifted to higher frequencies. However this is not the case, in particu-

lar during early infancy, and therefore articulatory gestures involving analogous anatomic and physiologic structures in the adult and the young infant will tend to result in acoustic outputs that are not linearly related to each other. To be sure, if the adult vocal tract were proportionally larger than the infant's, equivalent articulatory gestures would generally tend to result in proportional acoustic results[2]. Conversely, due to the lack of articulatory proportionality, when anatomic and physiologic equivalent actions are applied to each of the vocal tracts the adult's and the infant's vocal tracts will simply acquire different articulatory configurations. The fact that analogous articulatory gestures in young infants and adults lead to different acoustic consequences raises the question of the alleged phonetic equivalence between the infant's and adult's speech sound production. In fact it is not trivial to equate production data from infants with their supposed adult counterparts since phonetic equivalence will lead to different answers if addressed in articulatory, acoustic or auditory terms.

To gain some insight on this issue a crude articulatory model of the infant's vocal tract was used to calculate the resonance frequencies associated with different articulatory configurations. The model allows a virtual up-scaling of the infant's vocal tract model, to "match" a typical adult length, enabling the "infant's formant data" to be plotted directly on the adult formant domain. With this model the acoustic consequences of vocalizations while swinging the jaw to create a stereotypical opening and closure gesture could be estimated for between larynx and vocal tract length corresponding to an infant, a child and an adult speaker (Lacerda 2003). An important outcome of this model is that the vowel sounds produced by the "young infant" vocalizing while opening and closing the jaw are differentiated almost exclusively in terms of their F_1 values. Variation in F_2 appears only once the larynx length increases towards adult proportions (Robb & Cacace 1995). Equivalent opening and closing gestures in the adult introduce some variation in F_2, in particular when the jaw is wide

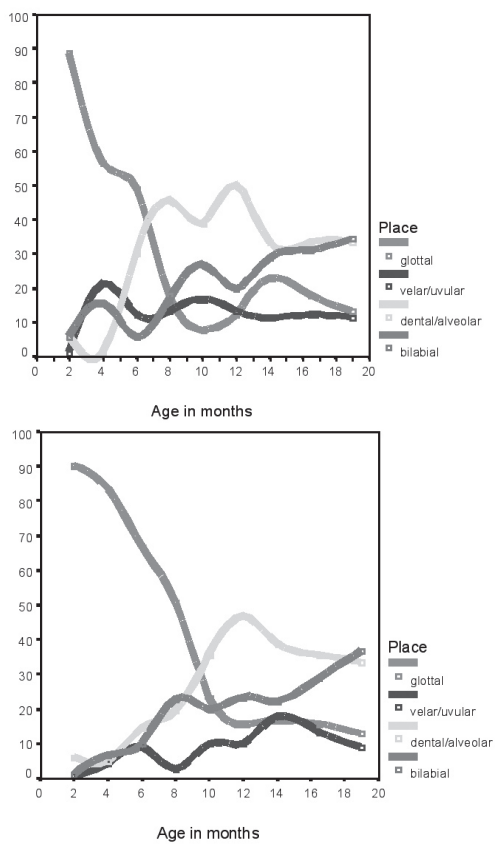
[2] Departures from this acoustic proportionality would however be observed for situations where the cross sectional area of the infant's proportional vocal tract would turn out to be under absolute critical values associated with turbulent or laminar flow through constrictions.

open because the ramus tends to push back and squeeze the pharynx. Since the infant's pharynx is much shorter, an infant producing the same swinging gesture will not generate appreciable variation in F_2. To be able to modulate F_2 in a similar way, the infant would have to pull the tongue body upwards toward the velum, probably by contracting the palatoglossus and the styloglossus muscles. For this configuration the infant's vocalization would have approximately the same formant structure as an adult [ɑ] vowel. To produce more front vowels, the infant would have to create a constriction in a vocal tract region at a distance about 70% full vocal tract length from the glottis, as predicted by the Acoustical Theory of Speech Production (Fant 1960; Stevens 1998).

A clear implication of the anatomical disproportion between infants and adults is that the early infant babbling sounds in general cannot be directly interpreted in adult articulatory phonetic terms. Of course acoustic-articulatory relations are even problematic in the adult case because the acoustic-articulatory mapping is not biunivocal, as demonstrated by the everyday examples of individual compensatory strategies and byte-block experiments (Gay, Lindblom & Lubker 1980, 1981; Lindblom, Lubker & Gay 1977; Lane, Denny, Guenther, Matthies, Menard, Perkell, Stockmann, Tiede, Vic & Zandipour 2005), but the problem is even more pertinent in the case of the infant's production of speech sounds (Menard, Schwartz & Boe 2004). Indeed, in addition to the non-proportional anatomic transformation, the infant, in contrast with the adult speaker, does not necessarily have underlying phonological targets, implying that infant vocalizations in general ought to be taken at the face-values of the acoustic output rather than interpreted in terms of phonetically motivated articulatory gestures. Disregarding for a moment experimental results suggesting that the young infant has the ability to imitate certain gestures (Meltzoff & Moore 1977, 1983) and even appears to be able to relate articulatory gestures with their acoustic outputs (Kuhl & Meltzoff 1982, 1984, 1988, 1996; Meltzoff & Borton 1979), the assumption here is that the infant is initially not even attempting to aim at an adult target sound. Thus in a situation of spontaneous babbling, where the infant is not presented with a "target" sound produced by a model, the phonetic content of the infant's productions must be strongly determined by

the infant's articulatory biases and preferences while the infant's productions (which often are rather diffuse in phonetic terms) will be perceived according to the expectations of the adult listener. As a consequence, the infant's vocalizations in a normal communicative setting are prone to be influenced from the very beginning by the adult's knowledge and expectation on the language. This type of interpretative circularity is indeed an integrate component of the very speech communication process and must be taken into account in studies of language development as well as in linguistic analyses.

Consider, for instance, the Swedish infants' preferences for different articulatory positions illustrated in figure 1, redrawn here after data from Roug, Landberg & Lundberg (1989).



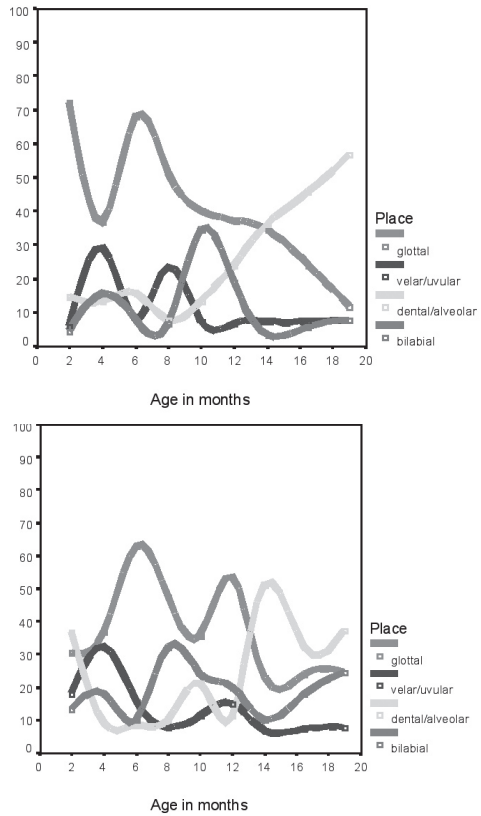


FIGURE 1 – Occurrence of place of different places of articulation in babbling produced by four Swedish infants. Original data from Roug et al. (1989).

In spite of the individual preferences, the overall trend of these data indicates a preference for velar and pharyngeal places of articulation during the first months of life. Pharyngeal places of articulation are typologically less frequent than places of articulation involving contact somewhere between the lips and the velum but they were nevertheless dominating in these infants' early babbling. But although the bias towards the pharyngeal places of articulation is likely to be a consequence of the vocal tract anatomy, where the young infant's extremely short pharynx leads to pharyngeal-like formant patterns when the infant's tongue actually makes contact with the velum or

palate, the infant's productions are actually perceived by the adult as being produced far back in the vocal tract, which translated in adult proportions means pharyngeal articulations. In a sense there is nothing wrong with this adult interpretation. It is correct from the acoustic perspective; the problem is that it is mapped into an adult vocal-tract that is not an up-scaled version of the infant's. Thus, because of this non-proportional mapping, acoustic or articulatory equivalence between sounds produced by the infant or the adult lead necessarily to different results and may lead to in overestimation of the backness of the consonantal sounds produced by the infant early in life (Menard *et al.* 2004).

Interactive constraints

All living systems interact one way or the other with their environment. At any given time the state of a living system is therefore the result of the system's history of interaction with its ecological context. Humans are no exception. They are simultaneously "victims" and actors in their ecological setting. They are the living result of long-term cross-interactions between internal genetic factors, the organism's life history and external ambient factors. Among the multivariable interactions observed in natural systems, language development is certainly an outstanding example of endogenous plasticity and context interactivity. As pointed out above, the infant's language acquisition process must be seen as an interactive process between the infant and its environment. In principle, in this mutual interaction perspective, both the infant and its environment are likely to be affected by the interaction. However, because the infant is born in a linguistic community that has already developed crystallized conventions for the use of sound symbols in communication, the newcomer has in fact little power against the community's organized structure. Obviously, the conventions used by the community are not themselves static but this aspect has very limited bearing on the general view of language acquisition. Not only do the users of the established language outnumber the infant but also the community language represents a coherent and ecologically integrated communication system, whose organization the individual newcomer can hardly threaten. In line with this, a crucial determinant of the infant's initial language development

must be given by the feedback that the community in general and the immediate caregivers, in particular, provide. Thus, if ambient speakers implicitly (and rightly) assume that the infant will eventually learn to master the community's language, they are likely to interpret the infant's utterances within the framework of their own language. In other words, adult speakers will tend to assign phonetic value to the infant's vocalizations, relating them to the pool of speech sounds used in the established ambient language and the adult will tend to reward and encourage the infant if some sort of match between the infant's productions and the adult expectations is detected.

To address this question, Lacerda & Ichijima (1995) investigated adult spontaneous interpretations of infant vocalizations. The infant vocalizations were a random selection of 96 babbled vowel-like utterances obtained from a longitudinal recording of two Japanese infants at 17, 34 and 78 weeks of age. The subjects were 12 Swedish students attending the second term of Phonetics[3] who were requested to estimate the tongue positions that the infant might have used to produce the utterances. The subjects indicated their estimates in a crude 5×5 matrix representing the tongue body's frontness and height parameters. In an attempt to simulate spontaneous adult responses in a natural adult-infant interaction situation, the students' judgements had to be given within a short time window to encourage responses on the basis of their first impressions. The answer forms consisted of a series of response matrices, one for each stimulus. The subjects' task was to mark in the matrices the cell that best represented tongue height and frontness for each presented utterance. If all the subjects would agree on a specific height and frontness level for a given stimulus, the overall matrix for that stimulus would contain 12 response units on the cell representing those coordinates and zero for all the other cells.

[3] Although students at this level are not strictly speaking naïve listeners, they represented a good enough balance between non-trained adults and the ability to express crude phonetic dimensions.

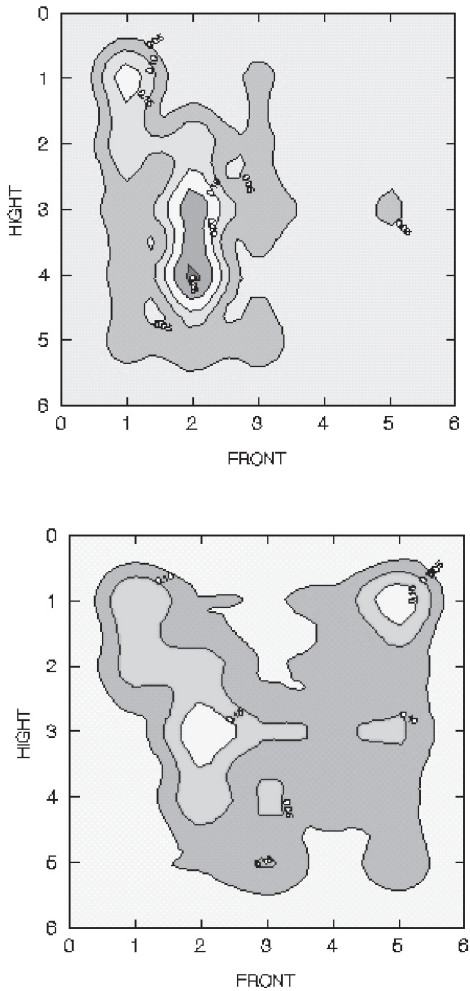


FIGURE 2 – The panels show level contours for agreement in adult spontaneous judgements of height and frontness of an infant's babbling. The left panel babbling produced at 17-34 weeks of age; right panel babbling produced at 78 weeks of age. High responses levels indicate that the adult subjects could agree upon the underlying articulatory gesture associated with the vowel-like utterance. The y-axis indicates the height dimension in an arbitrary scale, where 0 corresponds to a maximally closed vowel and 5 to a maximally open. The x-axis represents frontness, with 0 corresponding to extreme front and 5 to extreme back position of the tongue.

When the individual judgements were accumulated within each of the cells of the response matrices, two age-dependent patterns of frequently occurring responses emerged. For babbling utterances coming from the samples at 17 and 34 weeks of age, the judgements were scattered over a wide range along the vowel height dimension whereas the range of variation along the frontness dimension was rather limited. In contrast, the judgements of the samples from the 78 week-olds displayed an expanded range of variation both along the frontness and the height dimensions. In other words, whereas earlier babbling samples were interpreted as reflecting the use of mainly the height dimension, the responses to the later babbling suggest a tendency towards adult-like expansion of the vowel space. These results are illustrated in the two panels of figure 2.

Admittedly, this type of results depends both on the acoustic nature of the infant utterances and on the adult's auditory interpretations. In fact the results might reflect an adult judgement bias, some kind of vocalization preference from the infant or a mix of these biases. One possibility would be that the infant actually produces a variety of vowel-like utterances uniformly scanning the entire domain of the available articulatory space. In this case, the pattern of the results from earlier babbling would reflect the adults' inability to consistently estimate the frontness dimension. This would be an adult bias that might be referred to as a sort of "phonological filter" that the adult developed as a result of language experience. Another possibility would be that adults actually could pinpoint the infant's articulatory gestures but that the young infant simply does not produce vowel-like sounds that are differentiated along the frontness dimension.

Whatever the underlying reason for the observed response patterns, they reflect an actual interactive situation where adults and infants meet each other, linked by phonetics. Indeed, the adult and the infant are intimately connected with each other in the language communication process and the actual causes of the observed response patterns cannot be disentangled without the independent information provided by general theoretical models. At this point an acoustic-articulatory model of the infant vocal tract is a useful research tool because it provides some insight on the potential acoustic domain of the infant's articulatory gestures. Interestingly, as implied by the

acoustic-articulatory considerations addressed in the former section, the infant's short pharyngeal length clearly curtails acoustic variance in F_2, which is the acoustic correlate of the frontness dimension. Indeed, the notion that young infants might be mainly exploring the height dimension of the vowel space was corroborated in a follow-up listening test where a group of four trained phoneticians was asked to make narrow phonetic transcriptions of the babbling materials that had been evaluated by the students. The phoneticians carried out their task in an individual self-paced fashion. They were allowed to listen repeatedly to the stimuli, without response-time constraints, and requested to rely on the IPA symbols, with any necessary diacritics, to characterize the vowel-like utterances as well as possible. However, contrary to the typical procedures used in phonetic transcriptions of babbling, these phoneticians were not allowed to reach consensus decisions nor were they aware that their colleagues had been assigned the same task. The phoneticians' data were subsequently mapped onto the 5×5 matrices that had been used by the students by imposing the 5×5 matrix on the IPA vowel chart, assuming that the "corners" of the vowel quadrilateral would fit on the corners of the domain defined by the matrix. The results of the phoneticians' IPA transcriptions and the subsequent mapping procedure essentially corroborated the overall pattern of the students. The consistency between the judgements by the students and by the trained phoneticians discloses a provocative agreement between the two groups' spontaneous interpretations of infant babbling. Once again, it is not possible to resolve the issue of whether this agreement is a consequence of a strong phonological filter or of acoustic-articulatory constraints in the infant's sound production. The data can also be interpreted as suggesting that the infant's utterances are articulatorily so undifferentiated that only main aspects of their production can consistently be agreed upon by a panel of listeners. Apparently, it is only the height dimension that is differentiated enough to generate consistent variance in the adult estimates. To be sure, it cannot be excluded that all that the young infant is doing during these vocalizations is to open and close the mouth keeping the tongue essentially locked to the lower jaw (Davis & Lindblom 2001; MacNeilage & Davis 2000a, 2000b; Davis & MacNeilage 1995). This account is supported by the theoretical

evidence from the previous acoustic-articulatory model using short pharyngeal length (Lacerda 2003) and suggests that the spontaneous adult judgements of the infant's babbling reflect rather accurately the major features of the infant's actual phonetic output.

In summary, the acoustic-articulatory data, the adult perception data and the acoustic-articulatory model, all provide consistent support to the notion that the infant initially uses mainly the high-low dimension of the vowel space. The full exploration of the vowel space, including the use of the front-back dimension, comes at a later age.

Clearly, the interactive constraints affect much more than just vowel perception. As stated above, the infant and the adult interacting with each other create a particular setting of mutual adjustments to reach the overall common goal of maintaining speech communication (cf. Jakobson's (1968) phatic function of speech communication), where adults modify their speaking styles in response to their own assumptions on the infant's communication needs and intentions. These speech and language adjustments involve modifications at all linguistic levels, like frequent prosodic and lexical repetitions, expanded intonation contours and longer pauses between utterances. In addition to these suprasegmental modifications, there is evidence of more detailed modifications observed at the segmental level. For instance, when addressing 3-month-olds adults seem to expand their vowel space in semantic morphemes (Van der Weijer 1999), as demonstrated by cross-linguistic data from mothers speaking American English, Russian and Swedish (Kuhl, Andruski, Chistovich, Chistovich, Kozhevnikova, Ryskina, Stolyarova, Sundberg & Lacerda 1997). The adult production of consonants is also affected in the adult-infant communication. The voice/voiceless distinction expressed by VOT (Voice Onset Time) seems to be systematically changed as a function of the infant's age while preserving the main pattern of durational relations that occur in the adult language[4]. In infant-directed speech (IDS) to 3-month-olds, the VOT is significantly shorter than in corresponding utterances in adult-directed speech (ADS), leading to an increased overlap of the voiced and voiceless stops (Sundberg & Lacerda 1999).

[4] This is true at least for Swedish, where it has been observed that the complementary quantity distinctions were preserved in IDS to 3-month-olds.

In contrast, as suggested by preliminary results (Sundberg & Lacerda, under review), VOT in IDS to 12-month-olds seems to provide a higher differentiation between voiced and voiceless stops than is typically observed in ADS. Within the framework of the communication setting, such adult phonetic and linguistic modifications can be interpreted as an indication that the adult adjusts to the communicative demands created by the situation. Putting together the suprasegmental, the VOT and the vowel formant data a picture of adjustments to the infant needs emerges. Exploring the significance of variations in F_0, the adult may modulate the infant's attention during the adult-infant interaction in order to keep the infant at an adequate state of alertness. When the adult-infant communication link is established, the adult seems to intuitively offer speech containing clear examples of language-specific vowel qualities embedded in a rich emotional content. Because F_0 modulation is essentially conveyed by the vowels, their phonetic enhancement is quite natural, while VOT distinctions may have less phonetic specification at this stage. Although the perceptual relevance vowel enhancement and VOT reduction have not been experimentally assessed from the infant's perspective, a plausible interpretation is that the adult intuitively guides the infant towards language.

One possibility is that the adult opportunistically explores the infant's general auditory preference for F_0 modulations. The adult adds some excitement to the infant perception by spicing the F_0 contours with "crispy" consonants that by contrast enhance the vowel segments (Stern, Spieker, Barnett & MacKain 1983). This patterning of the speech signal is likely to draw the infant's attention to the vocalic segments. A further indication of the adult's adaptation to the perceived demands of the infant is provided by the change in phonetic strategy that is observed when adults address infants at about 12 months of age. Triggered by the clear signs of language development provided by the first word productions and proto-conversations, the adult introduces phonetic explicitness also to the stop-consonants, as suggested by data being processed at Stockholm University's Phonetics Laboratory. Obviously, in this scenario, the communication process between the adult and the infant cannot be equated to an adult-to-adult conversation. Rather, the adult appears to use language as an efficient and interactive playing tool, where speech sounds are the toys themselves. Through this

adaptive use of the language, the adult explicitly provides the infant with proper linguistic information wrapped in the playful and efficient setting of the adult-infant interaction. The infant is spontaneously guided by the adult's clear cues towards the general conventions of the speech communication act, like phonetic markers of turn-taking such as F_0 declination, final lengthening and pause duration (Bruce 1982).

Yet another conspicuous aspect of interaction constraints is the frequent use of repetitive patterns in infant-directed speech. As evidenced by a number of cross-language studies, adults interacting with infants tend to repeat target words and phrases as well as intonation patterns (Papousek & Hwang 1991; Papousek & Papousek 1989; Fernald, Taeschner, Dunn, Papousek, de Boysson-Bardies & Fukui 1989). Stockholm University's experimental data on mother-infant interactions clearly indicate that adults are extremely persistent in using, for example, words and phrases in a highly repetitive manner when addressing young infants. In the context of memory processes and brain plasticity this is likely to be a powerful strategy, as predicted by the language acquisition model that generates emergent patterns as a direct consequence of the interaction between memory decay and exposure frequency (Lacerda, Klintfors, Gustavsson, Lagerkvist, Marklund & Sundberg 2004).

To take the infant's perspective, it is mandatory to discuss the impact that the adult's phonetic modifications and repetition patterns may have for the infant.

Perceptual constraints

The newborn infant is exposed to a wide variety of information from the onset of its post-natal life. The infant's contact with its external environment is mediated by the information from all the sensory input channels, in continuous interaction with the infant's endogenous system. Clearly, the infant's initial perception of the world has to rely on the range of the physical dimensions in the environment that the infant's sensory system can represent[5]. From the current point of view,

[5] An organism's sensory system can be seen as a filter attending to a limited range within each of the environment's physical dimensions. The human visual system, for

because the infant is not assumed to be endowed with specific language-learning capabilities, the infant's perceptual capabilities at birth can be expected to be important determinants of the infant's development. The infant's auditory capability, in particular, is likely to be an important, though not sufficient, pre-requisite for the development of the speech communication ability, so let us try to get a picture of the infant's initial auditory capacity by relating it to the adult's.

From the onset of its post-natal life, the infant seems to have an overall frequency response curve that is essentially similar to the adult's, though shifted upwards by about 10 dB (Werner 1992). Also the quality factors of the infant's tuning curves are comparable to the adult's, at least in the low frequency region up to about 1 kHz. Given this resemblance, the infant's and the adult's auditory systems may be expected to mediate similar sensory representations of the speech signals, implying that differences in behavioural response patterns to speech stimuli may be attributed to higher-level integrative factors rather than peripheral psychoacoustic constraints.

Infant speech discrimination studies involving isolated speech sounds typically demonstrate that young infants are able to discriminate a wide variety of contrasts, virtually all the speech sound contrasts that they have been tested with. In fact, even 4-days old infants have been shown to discriminate between CV bursts as short as 20 ms (Bertoncini, Bijeljac-Babic, Blumstein & Mehler 1987), suggesting that the newborn infant is equipped with the necessary processing mechanisms to differentiate between bilabial, dentoalveolar and velar stop consonants. These and similar results from speech discrimination experiments (Eimas 1974) with young infants demonstrate that there is enough acoustic information to discriminate the stimuli. This is likely to be a pre-requisite for linguistic development but discrimination ability, by itself, is clearly not enough. In fact, discrimination alone

---

instance, is adapted to represent the range of electromagnetic radiation with frequencies between the infrared and ultra-violet and the auditory system reacts to changes in the atmospheric pressure falling within a limited range of amplitude and frequency whereas other species pick up other ranges of these physical dimensions. Given the differences in the representation ranges, the "reality" available to the different species is likely to be different too.

is likely to generate a non-functional overcapacity of separating sounds on the basis of their acoustic details alone. Linguistically relevant categories explore similarities among speech sounds that go beyond the immediate acoustic characteristics, as it is the case of allophonic variation or the same vowel uttered by female or male speakers. Such sounds are easily discriminable on pure acoustic basis but are obviously linguistically equated by competent speakers. Thus, a relevant question that has to be addressed by the model concerns the processes of early handling of phonetic variance underlying the formation of linguistically equivalent classes and the question of how the infant's initial discrimination ability relates to potential initial structure in the infant's perceptual organisation must be addressed. Parallel to the main trend of the experimental results pointing to a general ability to discriminate among speech sound contrasts, a remarkable asymmetry in the discrimination of vowel contrasts was observed in Stockholm University's Phonetics Laboratory. Experiments addressing the young infants' ability to discriminate [ɑ] *vs.* [a] and [ɑ] *vs.* [^], indicated that the latter contrast was more consistently discriminated than the former (Lacerda 1992a, 1992b; Lacerda & Sundberg 1996). The stimuli were synthetic vowels that differed only in their F_1 or F_2 values[6], reflecting a contrast in sonority (i.e. vowel height dimension, F_1) or in chromaticity (i.e. vowel frontness dimension, F_2). The contrasts were conveyed by equal shifts in F_1 or F_2, expressed in Bark. The results were obtained from two different age ranges and subject groups, using age-adequate techniques. The younger infants, around 3 months of age, were tested using the High Amplitude Sucking technique (Eimas, Siqueland, Jusczyk & Vigorito 1971) and the older subjects, 6 to 7 months old, were tested with the Head-Turn technique (Kuhl 1985). In both cases the outcome was that discrimination of the contrasts involving differences in F_1 dimension was more successful than for the corresponding contrasts conveyed by F_2 differences, in spite of the fact that equal steps in Bark

[6] F_1 and F_2 refer to the first and second resonance frequencies of the vocal tract, i.e. formants. The first two formants are usually enough to specify the main characteristics of a vowel's phonetic quality.

actually mean larger differences in F_2 than in F_1 if expressed in Hz. In principle this discrimination advantage for the F_1 contrasts might be attributed to the concomitant intensity differences that are associated with changes in F_1 frequency but intensity was also strictly controlled in follow-up experiments using a parallel synthesis technique (where overall intensity is not dependent on formant frequency) and yet the F_1 advantage in discrimination performance persisted. Thus, these experiments suggest that the infant's perceptual space for vowels is asymmetric in terms of height and frontness contrasts, with a positive discrimination bias towards height.

Putting the infant in its global context of production, perception and interaction, there seems to be an inescapable pattern of asymmetry that tends to enhance vowel contrasts along the height dimension. As stated above, data from infant vowel production, adult-infant interaction and infant vowel perception, all converge towards a pattern of dominance of the height dimension in early language acquisition. These results are also consistent with typological data from natural vowel systems (Maddieson & Emmorey 1985; Maddieson 1980; Liljencrants & Lindblom 1972). To the extent that infant speech perception and the adult's interpretation of babbling provide an indication of the biases underlying language development in general, vowel height may be expected to play a dominant role in the organization of vowel systems. In fact, this is in good agreement with the typological data showing that vowel height seems to be the first single explored dimension in vowel systems of increasing complexity whereas frontness contrasts usually are accompanied by rounding gestures, as if to underline the frontness distinction (Liljencrants & Lindblom 1972).

In summary, the overall message provided by the speech perception experiments with infants is indeed compatible with the notion that the infant starts off with a general auditory process that gains linguistic content in the course of the language acquisition process. Young infants are reportedly good at discriminating speech sound contrasts but their discrimination can largely be accounted for by sensitivity to acoustic differences per se, not necessarily liked to underlying linguistic strategies. A large body of speech perception studies in which infants were tested on discrimination of both native and non-native speech sound contrasts indicates a progressive attention focus

towards sound contrasts that are relevant in the ambient language. With respect to vowel perception, for instance, it was observed that 6-month-old infants tend to display a vowel discrimination behaviour that seems to have been influenced by their exposure to the ambient language (Kuhl, Williams, Lacerda, Stevens & Lindblom 1992). Infants at this age show a higher tolerance to allophonic variation for the vowels occurring in their native language than for non-native vowels, a phenomenon often referred to as the "perceptual magnet effect" (Iverson & Kuhl 1995, 2000; Kuhl 1991; Lotto, Kluender & Holt 1998). This process seems to be accompanied by increasing attention focus on the ambient language and, by about 10 months of age, infants may no longer be able to discriminate foreign vowel contrasts (Polka & Werker 1994). Also the perception of consonantal contrasts appears to follow a similar developmental path, although the development is shifted upwards in age. Whereas by about 6 to 7 months of age no particular differences in the discrimination ability for native and non-native consonantal contrasts have been observed, at 12 months of age infants seem to be more focused on the native than on the non-native consonant contrasts (Werker & Tees 1983, 1992; Best, McRoberts & Goodell 2001; Werker & Logan 1985; Tees & Werker 1984; Werker, Gilbert, Humphrey & Tees 1981). Taken together, these experimental results seem to be that exposure to the ambient language shifts the infant's focus from a general to more a differentiated and language-bound discrimination ability (Polka & Werker 1994; Polka & Bohn 2003).

Much of the speech discrimination studies assessing the infant's early capabilities have been carried out using isolated speech sounds. Isolated speech sounds and their underlying phonemic representations, as portrayed in linguistic theories, are useful for logical and formal descriptions of language. Nevertheless the phonemic concept is not obviously connected to the speech sounds that supposedly materialize it. Phonemes are idealizations that capture the essential contrastive function in language and are therefore not immediately available to the young language learner. What the infant is exposed to are strings of interwoven speech sounds, with all the concomitant co-articulation effects and non-canonical aspects affecting all levels of connected natural speech. Isolated speech sounds are rare in natural interaction,

raising problems to the linguistic interpretation of many infant speech discrimination experiments, in particular when single speech sound tokens are taken to represent a phonemic category. In particular, the lack of variance in the stimuli presented to the infant is likely to severely limit the ecological validity of the results but the linguistically relevant issue is to try to find out how the infant, handling natural variance, nevertheless homes in on an adult-compatible linguistic representation. The challenge is to account for language acquisition building on the fuzziness and variability that are characteristic of the infant's natural environment.

About a decade ago, infant speech perception studies started to address this issue using a variety of experimental approaches. For example, in an attempt to assess the significance of repetitive structures embedded in a continuous speech signal, Saffran and colleagues (Saffran, Aslin & Newport 1996) presented 8-month old infants with sequences of concatenated CV-sequences drawn from a set of four basic CV-sequences. The Bayesian conditional probability of the concatenated CV-sequences was manipulated to ensure that certain CV pairs would occur with higher probability than others. This was done to reflect natural language structure, in which syllable sequences within words tend to have higher transitional probabilities than syllable sequences across words. After a 2-minute exposure to this type of material, the infants showed a significant preference for the pseudo-words formed by syllables of high transitional probability, suggesting that they had been able to pick up implicit statistical properties of the speech material (Saffran & Thiessen 2003; Saffran 2002, 2003; Seidenberg, MacDonald & Saffran 2002). Also a number of studies carried out by the late Peter Jusczyk and colleagues suggest that infants are sensitive to high frequency words in their ambient language. The group reported, for instance, that four-month-old infants were sensitive to the high frequency exposure to their own names (Mandel & Jusczyk 1996; Mandel, Kemler Nelson & Jusczyk 1996) and also that nine-month-olds, in contrast with 7½-month-olds, were able to pick up high frequency words from the speech stream of a story telling (Johnson & Jusczyk 2001; Mattys & Jusczyk 2001; Houston, Jusczyk, Kuijpers, Coolen & Cutler 2000; Nazzi, Jusczyk & Johnson 2000; Jusczyk 1999).

These experiments represent different scenarios of language exposure. In the experimental set up of Saffran *et al.* (1996) only the acoustic information provided by synthetic utterances was available to the infants, who nevertheless could use the statistical regularities to structure the continuous sound streams in spite of the limited exposure and sparse information load of the signal. Jusczyk's work provides a closer match to a natural language acquisition setting since, in addition to the audio signal, the infants also had access to picture books providing visual support to the story they were exposed to. In comparison with Saffran and colleagues' set up, Mandel and Jusczyk's experiment clearly offered a much richer linguistic environment due to the variance in the speech material that the infants were exposed to (Mandel & Jusczyk 1996). By itself, the audio signal available to Mandel and Jusczyk's subjects during the training sessions is not nearly as explicit as in Saffran's set up. However, the total amount of exposure (with daily exposure sessions carried out for about two weeks) was far more extensive in Jusczyk's experiment. In addition to this longer exposure, the infants were also encouraged to look at a picture book, which may have been a critical component for the positive outcome of the experiment. Indeed, in line with the ideas expressed in this article, linguistic meaning emerges from the co-varying multi-sensory information available during exposure – in Mandel and Jusczyk's case, the naturalistic speech co-varying with the visual information.

To study the significance of co-varying multi-sensory information a series of experiments designed to create learning situations from controlled multi-sensorial information was recently started. To assess the infant's ability to link acoustic and visual information in a linguistically relevant way a variant of the visual preference technique has been used. The preliminary results suggest that 8-month-old infants are able to establish linguistic categories, such as nouns, from exposure to variable but consistent audio-visual information (Lacerda, Sundberg, Klintfors & Gustavsson, forthcoming).

Modelling the infant in an ECOLOGICAL setting

Background and overview of an Ecological Theory of Language Acquisition

This section presents a general model of a system capable of learning linguistic referential functions from its exposure to multi-sensory information (Lacerda, Klintfors *et al*. 2004). By itself, the model has a wider and more abstract scope than what has traditionally been considered as language learning. Rather than focusing on the acoustic signal per se, as the main determinant of the acquisition of spoken language, it is here suggested that the language acquisition should be addressed as a particular case of a general process capturing relations between different sensory dimensions. In this view linguistic information is implicitly available in the infant's multi-sensory ecological context, and is derivable from the implicit relationships between auditory sensory information and other sensory information reaching the infant. Early language acquisition becomes therefore a particular case of a more general process of detection of contingencies available in the sensory representation space.

Two initial assumptions are made in this model: one is that there is multi-sensory information available to the system and the other is that the system has a general capacity of storing the incoming sensory information but this latter assumption does not mean that the system will permanently store information or even all the incoming information.

In the case of the infant, this input is thought to consist of all the visual, auditory, olfactory, gustatory, tactile as well as kinaesthetic information. Maintaining life requires continuous interaction between the living system and its environment, e.g. the complex organisms' basic life-supporting functions like breathing and eating. To succeed, the biological system must be able to handle information available in its environment and use it to acquire the necessary life-supporting resources. In this sense, information obviously refers to a vast range of physical and chemical properties of the organism's immediate environment (like harshness and structure of surfaces with which the organism makes contact, chemical properties of the environment, pres-

sure gradients, electromagnetic radiation, etc.) as well as information conveyed by potential relationships between these elements, like the underlying relation between sounds of speech and the visual properties of the objects that the speech signal might refer to. To be sure, living organisms tend to specialize on processing quite a limited range of the potentially available physical and biochemical diversity available in their environments, i.e., different systems specialize in more or less different segments of the available environment, exploring the potential advantages of focusing on specific ecological niches. Numerous examples, like bats exploring echolocation to survive in the ecological niche left open by their daylight competitors, or the specialized bugs, like the Agonum quadripunctatum or the Denticollis borealis (Nylin 2006), that only emerge in the special environment created by the aftermath of forest fires. In these very general terms, the external environment is represented by changes in the system itself, changes that may in turn be responses and generate further interaction with the environment, and that a species evolutionary history has prompted the organism to select a segment of.

Also the infant must be regarded as biological system integrated and interacting with its environment. The infant's environment is represented by sensory information conveyed by the sensory transducers interfacing with the environment, i.e. the changes that the environment variables induce in the specialized sensory transducers. To be sure, the infant's sensory system's response to the environment variables is not time-independent. Typically, as the infant learns and develops, its responses will change not only of the exposure to external stimuli but also as a consequence of the very changes in internal, global variables that are induced by that exposure. Given that the infant is continuously exposed to parallel sensory input simultaneously available from different sensory modalities, there the potential for combining this multimodal information into different layers of interaction. The hippocampus, for instance, is thought to be a structure capable of integrating different sources of sensory information in implicit memory representations (Degonda, Mondadori, Bosshardt, Schmidt, Boesiger, Nitsch, Hock & Henke 2005). However the perspective in this article is more functional and somewhat abstract. Sensory inputs are seen as n-dimensional representations of the

external variables, and that such representations may be continuously passed to a general-purpose memory that will retain at least part of that information. Also the concept of memory is very general. It is a general-purpose memory space on which continuous sensory input is represented by changes in the detail-state of the organism that can be modelled as Hebbian learning (Munakata & Pfaffly 2004). The sensory input is continuously mapped in this memory space, in a purposeless and automatic way but representations that are not maintained tend to fade out with time.

The infant's ecological settings

In its ecologic setting, the infant is inevitably exposed to huge amounts of sensory input that at first sight may appear to be unstructured. At a closer look, however, the infant's immediate world is indeed highly structured in the sense that the status of the parallel sensory inputs implicitly reflects the aspects of the structure of objects and events that are being perceived (Gibson & Pick 2000). Language is part of this and the infant's early exposure to it is indeed also highly context-bound and structured. In typical infant-adult interactions, the speech used by the adult is attention catching, circumstantial and repetitive, making this sort of speech attractive for the infant, context bound and easily related to external objects and actors. Under these circumstances, constantly storing general sensory input under the exposure to the statistic regularities embedded in the external world must lead to the emergence of the underlying structure because memory decay effectively filters out infrequent exposures. So far, this formulation may seem rather close to classic Skinnerian reasoning and exposed to the classical criticism calling on the "poverty of the stimulus" argument. However, taking into account the very variance of the input, it is possible to tip the argument over its head and instead used as a resource for language learning.

From a global and long-term perspective language obviously builds on regularities at many levels, reflecting the conventionalized communication system with which individuals can share references to objects and actions in their common external, as well as internal and imaginary, worlds. On a short-term basis, the statistical stability of those regularities may not always be apparent, especially for the

newborn. However, lacking long-term experience with language and knowledge of the world, the newborn is happily unaware of the potential problems that linguists assume infants to have. In other words, early language acquisition is not seen as a problem since the infant presumably does not have a teleological perspective. The infant may very well converge to adult language as a result of a parsimonious and "low-key tinkering process" that, under the pressure of local contingencies, leads to disclosure of the implicit structure of language. In fact, just because language is always used in a context – particularly at the early stages of infant-adult interactions, where language tends the focus on the proximate context – there are plenty of systematic relations between the acoustic manifestations of language and their referents.

Sketching an ecological setting of early language acquisition

General background

The present sketch of the language acquisition model assumes that sensory input is represented by values in an n-dimensional space, where each dimension arbitrarily corresponds to a single sensory dimension. Obviously, in this generic model, sensory dimensions like auditory input may themselves be further represented by m-dimensional spaces to accommodate different relevant auditory dimensions, but for this sketch a general and principled description will be sufficient. These external stimuli are mapped by the sensory system into the internal representation space. The mapping process does not involve any explicit interpretation of the input stimulus. It is viewed essentially as a sensory map affected by sensory limitations and the stochastic representation noise inherent to the system. This means that rather than directly converting the input into coordinates on the representation space, the sensory system adds a certain amount of noise to the input, reflecting the stochastic nature of neuronal activity. As a consequence of the added noise physically identical external stimuli will tend to be mapped onto overlapping activity distributions centred at neighbouring rather than identical coordinates on the representation space. At this stage it is assumed that the neuronal noise is uncorrelated with the input and that it has zero mean value, an assumption that essentially means that the noise's long-term effect can be disregarded.

To calibrate the proposed model, taking into consideration its ecological context it is convenient to start out with an estimate of the magnitude and range of variation available in the model's n-dimensional sensory space. To create a manageable model, a 2-dimensional space, supposedly representing the auditory and visual (A-V) sensory inputs[7], was chosen here. Activity on a location of the representational space corresponding to these two dimensions is a manifestation of co-occurrences of rather specific auditory and visual stimulations that are stochastically associated with the location's coordinates. Uncorrelated occurrences of auditory and visual sensory inputs will tend to scatter the corresponding representation activities. As a consequence, uncorrelated or random A-V activity tends not to build up at specific locations and does not contribute to structuring the representation space because activity will be smeared over a wide region. Looking at the sensory input from the perspective of the representational space, this lack of heightened localized activity means that there is no systematic relation between the input stimuli. But in the context of early language acquisition, uncorrelated A-V representations are not much more than a simple academic abstraction. Because language is used in a coherent way with its referents, ecologically relevant language acquisition settings are much more likely to provide correlated audio-visual information than not. In terms of the model, this will tend to raise the activity level in specific locations in the representation space. Put in these terms, a crucial question concerns the probability of hitting approximately the same location in the representation space, when the sensory input is mediating random, uncorrelated events.

However the notion of underlying sensory contingencies is not new. Behaviourists have attempted to account for learning in terms of reinforced stimulus-response correlations and been confronted with the "poverty of stimulus" argument. Therefore it is convenient to take a look at the dimensionality of this A-V search space in order

[7] This implies a linearization of the auditory and visual components, meaning that each possible auditory spectrum is assigned a unique point along the "auditory axis" and a similar representation for the visual stimuli, a transformation that indeed does not really capture the natural dimensionality of the auditory and visual spaces.

to estimate of the number of possible discriminable events in the auditory space that may be created by all possible combinations of discriminable intensity levels and frequency bands. To simplify the calculations, aspects like lateral suppression and masking effects that might affect the ability to discriminate between some of the theoretical sounds in this set are not taken into account. In reality, those effects would slightly reduce the number of distinct sounds but that is largely compensated by the rather conservative assumptions on frequency and intensity discriminable steps. On the intensity dimension, it is assumed that level differences of at least 5 dB can be discriminated which yields approximately 17 intensity steps along the intensity dimension throughout the audible spectrum, taking into account the frequency dependency of the hearing threshold. The frequency range between 100 Hz and 4000 Hz is assumed to be discriminable also in 17 steps, corresponding approximately to 17 Bark scale intervals. It is assumed that the linguistically relevant sounds can be represented by their crude energy distribution along these 17 Bark bands, a rather conservative frequency range from which speech sounds like many fricatives tend to be excluded.

Viewing the auditory stimuli as instances of 17-band spectra with 17 possible intensity steps per band leads to an estimate of 17^{17} ($\approx 10^{20}$) potentially discriminable spectral contours. This is, of course, a crude estimate whose main goal is to map possible acoustic stimuli onto a one-dimensional linear representation but it gives nevertheless a feeling for the range of possible spectral variation of relevant acoustic stimuli. An important consequence of estimating the domain of acoustic variation is that it gives an indication of the likelihood of hitting a given point in this space under a plausible time-window. 10^{20} is an amazingly large number. To get a feeling for its magnitude it may be imagined that each of those possible 10^{20} spectra is represented by a cell in a human body. With this analogy, it would be necessary to have about 10 million individuals, weighting 100 kg each, in order to obtain the 10^{20} cells corresponding to this crude estimate of the acoustic search space[8]. In such a space, hitting twice a given

[8] The calculation is based on the assumption that there are approximately 100 million cells per gram of biological tissue, relatively independent of the tissue (Werner 2006).

neighbourhood is an extremely significant event. Indeed, exploring the cell analogy, if one assumes that all those 10 million people pick up at random their travel destinations on the earth, chances are vanishing small that the same two people will happen to meet twice on those trips. Therefore, if one happens to meet the same neighbour of two trips, it is reasonable to view that second meeting as a very significant event, potentially suggesting that one is being followed by that neighbour. But even if it is not exactly the same neighbour that is met twice but just one of the nearest neighbours, the assumption of random trips suggests that even this is highly significant given the vanishingly small probability of that recurring random event. In other words, because of the extremely vast acoustic search space, recurrent events are extremely unlikely at random and therefore highly significant if they occur. Adding to the auditory space another sensory dimension, like the visual dimension, enhances even more the significance of recurrent correlated audio-visual events. Given the extremely low probabilities of repeated audio-visual events generated by random associations such repeated instances of correlated audio-visual events are even more unlikely than when the auditory dimension was considered alone. Thus, under ecologically relevant conditions an organism can "take for granted" these highly significant events and interpret them as systematic associations. In other words, events recurring a couple of times in this audio-visual space are so unlikely to happen at random that the organism can take them for granted. Of course such a generalization is not "safe" in absolute terms since the organism "jumps into conclusions" that are not clearly supported by the available data. However, the generalization risks involved in such "hasty" conclusions provide the organism with useful power in structuring its immediate environment and humans seem to be prone to jump into conclusions regarding potential relationships between different observed phenomena (Dawkins 1998). Such a generalization power has to be captured by the language acquisition model because it introduces a very important qualitative discontinuity in the interactive process between the organism and its environment. The over-generalization provides the organism with a hypothesis about the structure of the environment, a hypothesis that is taken for granted and used as an axiomatic building block in the ongoing interaction

with the organism's context. In other words, the organism generates a hypothesis about its relation with the environment and uses it as a framework to structure future interactions although it does not necessarily have full information about the potential correctness of that decision (Bechara, Damasio, Tranel & Damasio 1997)[9]. George Kelly's Theory of Personality and his psychology of personal "constructs" (Kelly 1963) also provides an interesting insight on the potential early mechanisms learning because it can be used to draw parallels between associations that individuals learn to view as unquestionable truths playing a fundamental role on their personality traits and the type of fundamental and unquestioned links that individuals learn to establish between the sounds of words and the objects (or actions) they refer to. Indeed, just like in the establishment of constructs that are accepted as self-evident truths, although they may be based on a few observations and sometimes wrong generalizations that become integrated in the individual's personality traits, early language acquisition is likely to be full of similarly obvious "constructs" between sound sequences and events that they are likely to refer to. Such "constructs" in the domain of language acquisition are also obvious and unquestioned sound-meaning relations that are simply assumed to be unquestionable truths and just because of that are useful as obvious building blocks in that emerging speech communication process.

The significance of co-occurrences

The domain of possible audio-visual variance is, as mentioned above, enormous and is therefore not practical to deal with directly. To simulate a realistic learning situation the model was scaled down for computational purposes, although reducing the domain of the audio-visual space impacts on the likelihood of hitting specific coordinates

[9] According to these authors frontal lobe injured patients lacked the ability to "jump into conclusions" until enough data were gathered to support a less risky conclusion. Although the particular mechanism of non-conscious somatic markers proposed by the authors has subsequently been questioned (Maia & McClelland 2004) and still is debated (Bechara, Damasio, Tranel & Damasio 2005), their study still illustrates the healthy subjects' propensity to integrate whatever information they have available in order to reach a quick and correct decision.

in that domain. Thus, to maintain realistic proportions between the domain and the likelihood of random hits, it is necessary to downscale the time-dimension as well. This means that reducing the number of cells in the audio-visual space must lead to a proportional reduction in the number of events processed by the model.

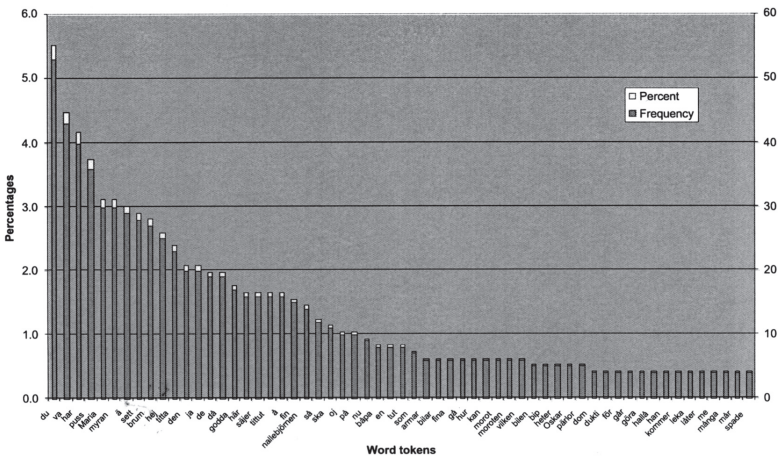


FIGURE 3 – Relative frequency and percentages of the top 80% forms used in a 12 -minute Infant-Directed Speech session.

These data come from a 12 minutes' mother-infant session to calibrate the model. The session was transcribed and the absolute and relative frequencies of word types and tokens calculated. The total number of word types throughout the session was 172, including play and hesitation utterances. The relative and absolute numbers of the top 80% word occurrences is shown in figure 3.

A typical aspect of this kind of token distributions is the high frequency use of a very limited number of these tokens. Of the 960 words used during the session, no more than 17 words (about 1.7%) account for 475 uses (i.e. nearly 50%), as listed in the table below. This distribution is somewhat similar to the observations of word use in adult speech communication but the proportions of item re-use are rather different. In a typical 3-minute sample of adult-directed speech recorded from a radio interview with a Swedish politician,

about 10% of the items accounted for the same 50% of word use, indicating therefore a much wider variation in the lexical items than in the case of infant-directed speech. Thus, the high repetition rate of a limited number of words embedded in slightly different linguistic and prosodic frames is likely to provide the infant with a highly efficient exposure to linguistic material from which the phonetic core of the tokens emerges.

A more detailed analysis of the material from the infant-directed speech, segmented in 60 s chunks, discloses an even more focused structure in the linguistic content of the infant directed speech. Indeed, during the first and the second minutes of the session the mother introduces a toy that she refers to as "myra" (eng *ant*) and the target word "myra" is repeated over and over again, as long as the infant's visual focus of attention is directed towards the toy. During the first minute the mother produces 122 word tokens, within which there are 10 occurrences of the target word, 8 occurrences of the attention modulating interjection "hej" (*hi*), 7 occurrences of the infant's name "Maria" and 5 occurrences of the words "du", "här" and "är" (*you, here* and *is/are*). In other words, only six word types account for about 1/3 of the total number of words during the first minute. Using the binomial distribution to estimate the probability of 10 occurrences of the same word type to be produced within this 60 s time window, under the assumption that all types would have equal probabilities of being produced, yields a probability of $p<0.000235$. Obviously, the probability of having one of the 47 types being repeated by chance 10 times among the 122 produced tokens is extremely low. Therefore the fact that such repetitions actually occur in infant-directed speech is an extremely significant event, whose importance can be taken for granted. Figure 4 displays the probabilities of different number of random occurrence of a given type for this first minute of the session. This pattern of significant repetitions is maintained throughout the session. For the second minute, for instance, the probability of random token repetitions associated with the target word is still as low as $p<0.000984$, for 28 types, 105 tokens and 11 repetitions of the most frequent word.

Word	Word translation	Frequency	Cumulative frequency	Percent	Cumulative percent
du	you	53	53	5.5%	5.5%
va	was, how	43	96	4.5%	10.0%
har	has	40	136	4.2%	14.2%
puss	kiss	36	172	3.8%	17.9%
Maria	Maria	30	202	3.1%	21.0%
myran	ant	30	232	3.1%	24.2%
är	is	29	261	3.0%	27.2%
sett	seen	28	289	2.9%	30.1%
brum	(play sound)	27	316	2.8%	32.9%
hej	hi	25	341	2.6%	35.5%
titta	look	23	364	2.4%	37.9%
den	the	20	384	2.1%	40.0%
ja	yes	20	404	2.1%	42.1%
de	they, them	19	423	2.0%	44.1%
då	then	19	442	2.0%	46.0%
goddag	hello	17	459	1.8%	47.8%
här	here	16	475	1.7%	49.5%
säjer	says	16	491	1.7%	51.1%

TABLE 1 – Tokens accounting for 50% of the occurrences in a 12-minute's Infant-Directed Speech session.

One might object to this kind of technical analysis as being a too simplistic exercise to account for a situation that is complex, very natural and familiar to many people. However, the value of the current approach is just that such rudimentary assumptions do expose essential aspects of the linguistic structure involved in the early language acquisition process. From a naturalistic point of view, repetitions occur as a consequence of the introduction of a new object that the mother intends to present to the infant. But from a human communication point of view referring to objects in the shared external world is exactly one of the essential functions of language, particularly at its early stages of development[10]. This is also the message from the probability estimates. If words were used at random it would be unlikely to observe the number of token repetitions that the infant

[10] It goes without saying that function of spoken language also includes other complex aspects such as behaviour regulation and information exchange.

is faced with in natural situations. Thus, it is the very obvious fact that language is not based on random use that underlines the significance of token repetitions. This is, of course, true both for adult-directed and for infant-speech but the number of repetitions occurring in infant-directed speech is more unlikely than that observed in adult-directed speech and this may very well trigger the infant's initial focus on the audio-visual contingencies that are thought to underlie the early language acquisition process. Indeed, these repetitions are not solely an auditory phenomenon and because they tend to be highly correlated with other sensory inputs, such as the visual and often tactile input, they build up salient multi-sensorial representations in the environment shared to the mother-infant dyad. In other words, recurrent audio-visual contingencies against the background of overall possible sensory variation are so unlikely to occur at random that a few repeated co-occurrences in this multi-sensory information domain can be treated as "sure indications" of the outside world's structure. However the current theoretical model does not propose that the infant will actually be seeking this type of correlations in order to learn its ambient language. In fact, it is assumed that detection of those important contingencies may initially be underlined by general purpose memory processes, provided the processes' time-windows are long enough to store some of these repeated events. Given the probability of repetitions in the early infant-directed speech, time windows of a few seconds will initially be enough to detect some of the recurrent audio-visual patterns in infant-directed speech.

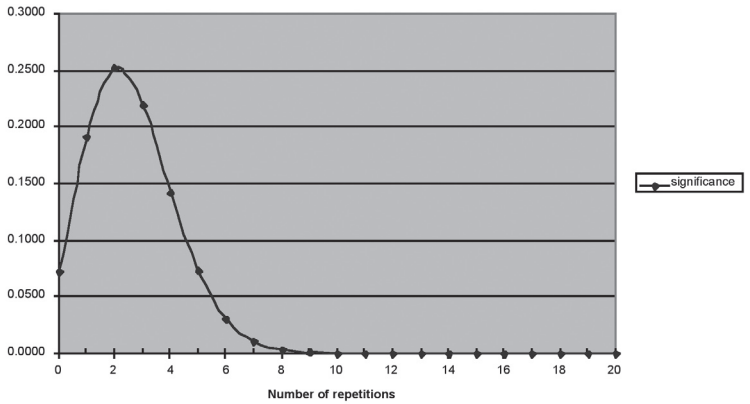


FIGURE 4 – Probability of random token repetitions considering a pool of 147 types and 122 trials.

While the audio-visual contingencies described above may trigger the early language acquisition process in infants, it is obvious that adults' speech does not consist of series of isolated words like the ones depicted above, but rather of chains of coarticulated speech sounds that the young language learner has to deal with in order to learn its ambient language. Dealing with chunks of speech sounds that are essentially continuous and not properly separated in word-like sequences raises a new challenge to the language acquisition process: How can words emerge from a continuous speech signal at the same time that the young language learner is assumed to lack linguistic insights?

An ecological theory of early language acquisition

What is the potential significance of repetitions in the speech material that the infant is exposed to? In order to have an ecologically relevant processing time frame, it is assumed that the young infant has a holistic auditory assessment of the speech signal. This means that in the absence of linguistic knowledge the infant will partition the incoming speech signal mainly on the basis of its acoustic level contour. In other words, the infant is assumed to perceive the signal as a non-segmented train of on and off sequences of sounds where

silences in between utterances are the only delimiters. To mimic this situation the utterances produced by one mother were represented as a series of character strings corresponding to the utterances' orthographic transcriptions but without considering any between-word boundaries that were not associated with an actual pause in the speech material (Lacerda, Klintforst *et al*. 2004; Lacerda, Marklund, Lagerkvist, Gustavsson, Klintfors & Sundberg 2004). The actual model input, corresponding to the first 60 s of mother-infant interaction was represented as shown below, where the commas simply represent pauses between successive utterances. The timeline of the original recording is kept in this sequence (figure 5).

hej, skavilekamedleksakidag, skavigöradet, ha, skavileka, tittahär, kommerduihågdenhärmyran, hardusett, titta, nukommermyran, dumpadumpadump, hej, hejsamyran, kommerduihågdenhärm yranmaria, ådenhärmyransomdublevalldelesräddförförragång en, ja, mariablevalldelesräddfördenhärmyran, ädemyranoskar, myranoskaräde, tittavafinmyranä, åvamångaarmarhanhar, hard usettvamångaarmarmaria, hejhejmaria, hallå, hallå, hallåhallå, hejhej, ojojsågladsäjerOskar, ämariagladidag, ämariagladidag, ha, hejmariasäjermyran

FIGURE 5 – Transcript of the input submitted to the model. The strings represent utterances separated by pauses. Words are concatenated if produced without pause in between.

Once a pause is detected, the on-going recording of the utterance is stopped and the utterance is stored unanalyzed in memory. The new utterance is then compared with previously stored utterances on a purely holistic basis, in order to find a possible match between the new utterance and those already stored. The search for a possible match is performed by considering two utterances at a time, one the newly stored utterance and the other an utterance drawn from the set already stored in memory. The shortest utterance in the pair is taken as a pattern reference and the other sequence is searched for a partial match with the pattern defined by the reference. If a match is found (figure 6) the common portion is assumed to be significant,

in accordance with the principles presented in the previous section, and gets its level of memory activity increased. The rationale here is that in the huge audio-visual space the likelihood of randomly finding two similar strings is vanishingly small, as discussed above, which means that repeated items can immediately be taken as good lexical candidates. The first loop of matches is displayed in the table below, where *n1* and *n2* refer to the items being compared and the pairs in curly brackets indicate the position of the matched elements on *n2*. The result of this process yields the activated items that are displayed in figure 7. These items become now part of the model's "lexical inventory".

n1=5	n2=2->{ {1, 9} }	n1=8	n2=4->{ {1, 2} }
n1=9	n2=6->{ {1, 5} }	n1=12	n2=1->{ {1, 3} }
n1=13	n2=12->{ {1, 3} }	n1=13	n2=1->{ {1, 3} }
n1=14	n2=7->{ {1, 23} }	n1=20	n2=9->{ {1, 5} }
n1=21	n2=4->{ {14, 15}, {17, 18} }	n1=22	n2=8->{ {1, 9} }
n1=22	n2=4->{ {1, 2} }	n1=23	n2=12->{ {1, 3}, {4, 6} }

FIGURE 6 – Example of identification of recurrent patterns: n1 and n2 indicate the positions in the transcript displayed in figure 5 of the utterances being compared. The numbers within curly brackets specify the substrings of n2 that match the content of n1.

Raw matches across utterances

skavileka, ha, titta, hej, hej, hej, kommerduihågdenhärmyran, titta, ha, hardusett, ha, hej, hej, ha, hallå, ha, hallå, hallå, ha, hejhej, hej, hej, ämariagladidag, ha, ha, ha, ha, ha, ha, ha, hej, hej

Independent lexical items:

ämariagladidag, ha, hallå, hardusett, hej, hejhej, kommerduihågdenhärmyran, skavileka, titta

FIGURE 7 – Raw matches and unique lexical items detected in the utterances of figure 5.

Items in this lexical inventory are now matched against both new and old (unanalyzed but stored in memory) input utterances. The effect

of this procedure is that the items in the lexical inventory can now be treated as (temporarily) acquired and therefore removed from the input utterances, thereby exposing the non-analyzed portions of the utterances. Applying this procedure to the original material leads to the non-analyzed chunks listed below. The procedure can be applied recursively to the list of non-analyzed chunks and in this case a new lexical item ("maria") can be derived.

medleksakidag, här, samyran, maria, vafinmyranä, åvamångaarmarnr, vamångaarmarmaria, maria, mariasäjermyran

FIGURE 8 – Additional lexical candidates identified in a second iteration, using the lexical items identified on the first interaction.

Further runs of the procedure on this limited material do not add to the number of stored lexical candidates. In a more realistic version of the model, the representations of both the audio input and of the lexical candidates have to be affected by memory decay but that is disregarded in the present example, as if the memory span would be long enough to represent the 60 s of input without appreciable degradation.

The procedure used for the simulated audio input can also be applied to the visual input. The visual component can be thought of as a series of images entering a short-term visual memory. This series of images is associated with a series of visual representations that will tend to overlap to the extent that there are common elements throughout the series[11]. Thus, the overlapping regions of the visual representations will implicitly yield information on recognizable visual patterns. And, just as in the case of the audio input, recurring visual patterns can be treated as significant elements, given the low likelihood of observing the same visual input would be repeated at random. This is illustrated in the following example.

[11] This is obviously only a principle description of how visual similarity may be detected. The computational processing required to establish matches between similar images has ultimately to be expressed in terms of convolutions, rotations, translations and scaling transformations, but at this point the focus is on the very general principles of visual processing, not with their explicit mathematical implementation.

Working on the audio-visual space

In this example it is assumed that the auditory and the visual dimensions can be linearized and arbitrarily represented by 100 steps along each of the dimensions[12]. A series of 5000 audio-visual possible events was created by combining uniform random distributions for both the audio and the visual dimensions. To mimic real-life audio-visual presentations of objects, the probability of three arbitrary but correlated audio-visual events was increased. Figure 9 illustrates this generated audio-visual space, where the three areas with heightened dot density reflect the referential use of the auditory information.

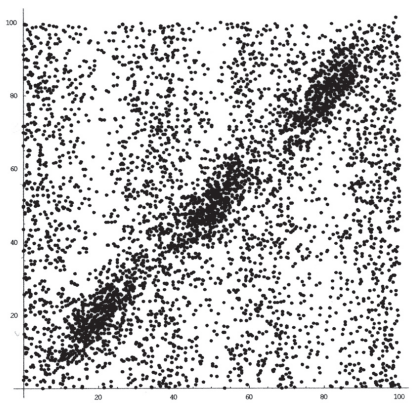


FIGURE 9 – Illustration of random A-V occurrences containing three recurrent A-V associations. One of the axes represents the visual dimension and the other represents the auditory dimension.

As the sensory input is mapped onto the representation space, it is affected by both sensory smearing and memory decay. Memory decay reflects a decreasing activity in the representation space, as a function of time. In this model memory is simply a volatile storage of mapped exemplars associated with the multi-sensory synchronic inputs[13]. New representations in this space add activity to the previously existing

[12] This is, of course, a microscopic domain in comparison with the actual physical world but it has the advantage of allowing the computations to be completed before retirement.

[13] Synchrony does not introduce a lack of generality since asynchronic but systematically related events can be mapped onto synchronic representations simply by introducing adequate time delays.

activity profile. In this representation space, overlapping neighbouring distributions convey an implicit similarity measure which makes sensory smearing crucial for the establishment of similarity patterns. The smearing value is thus critical for the model output. A too narrow smearing value leads to an overestimation of the number of categories in the representation space, whereas a too broad smearing spreads information in a meaningless way all over the representation space[14]. In this model smearing is viewed as two-dimensional Gaussian distributions centred on the incoming auditory-visual stimuli values. In the present example, based on 5000 simulated events, the memory decay was set so that a level of 30% of the initial activity would be reached after about 2000 time units (events). The sensory smearing was set to 30 sensory units in each of the two dimensions. Figure 10 displays the activity landscape after exposure to the 5000 stimuli shown above. The height of the hills in this landscape reflects both the cumulative effect of similar sensory inputs and proximity in time (memory decay implies that older representations tend to vanish).

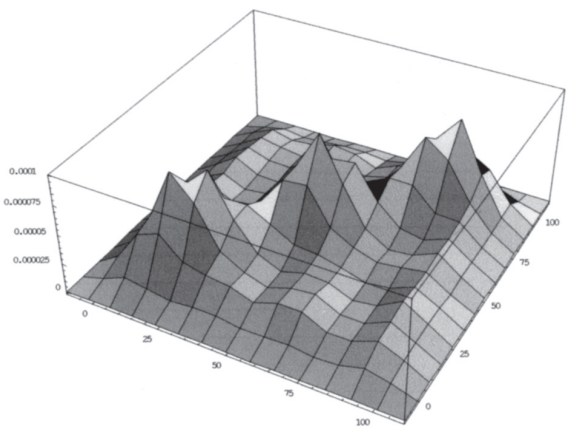


FIGURE 10 – Illustration of the memory activity created by the A-V stimuli simulated in figure 9, taking into account memory decay.

[14] The ability to organize information into equivalence classes is undoubtedly a central aspect of the organism's information handling capacities. Narrow discrimination ability may, by itself, become an obstacle to the information structuring process. Some of the language and learning disabilities observed in children may in fact be attributed to inability to disregard details in the incoming sensory information.

The model generates a representational space in which correlated external information tends to cluster into denser clouds, as a result of memory decay interacting with frequency of co-occurrence of the different sensory inputs. In the context of language acquisition, it may be expected that the initial regularities conveyed by the use of frequent expressions to refer to objects or actors in the infant's immediate neighbourhood will suffice to generate stable enough correlations between the recurrent sounds of the expressions and the referents they are associated to. Such stable representations form specific lexical candidates, both words and lexical phrases, that somehow correspond to events or objects in the linguistic scene. Obviously "words" do not necessarily correspond to established adult forms. They simply are relatively stable and non-analyzed sound sequences, like the ones picked up by the first model above, that can be associated with other sensory inputs. Once this holistic correspondence between sound sequences and their referents is discovered, it is possible for the infant to focus and explore similar regularities, thereby bringing potential structure to the speech signal that the infant is exposed to. In summary, the current model starts off with a crude associative process but it can rapidly evolve towards an active exploration and crystallization of linguistic regularities. Applied to actual language acquisition settings, the model suggests that the infant quickly becomes an active component of the language discovery process. Language acquisition starts by capitalizing on general sensory processing and memory mechanisms which tend to capture sensory regularities in the typical use of language. Recurrent consistency in the structure of the input stimuli leads to clustering in the representation space. The sensory relationships implicit in these clusters can subsequently be used in cognitive and explorative processes, that although not designed to aim at language, converge to it as a result of playful and internally rewarding actions (Locke 1996).

Conclusion

The question of how language communication in general and language development in particular have to rely on specialized language-learning mechanisms opened this article and it was argued

that both may perhaps be seen as unfolding from general biological and ecological principles. Resolving the issue of specialized versus general mechanism is virtually impossible, if based on empirical data alone. All living organisms are the result of their own cumulative ontogenetic history, reflecting the combined influences of biological, ecological and interactive components.

To pave the way for the view that language acquisition may largely be accounted for by general principles this article started with a short overview and discussion of possible evolutionary scenarios leading to the emergence of language communication. From the evolutionary perspective the ability to communicate using symbols does not seem to be exclusive for *Homo sapiens*. A number of other species do also represent relevant events of their ecological environment by using codes like sound, mimic or gestures but yet they fall short of human language because they seem to lack the capacity to use those symbols recursively (Hauser, Chomsky & Fitch 2002). In some sense, the process of early language acquisition can be seen as a modern revival of the evolutionary origins of language communication and from this perspective the study of early language acquisition may provide unique insights on how intelligent beings learn to pick up the regularities of symbolic representation available in their immediate ecological environment. There are, of course, very important differences between the modern situation of language acquisition and the evolution of language communication itself. One of those differences is the fact that infants now are born in an ecological context where language is already established and coherently and profusely used in their natural ambient. Another difference may be that *Homo sapiens* have evolved a propensity to attend spontaneously to relations between events and their sensory representations, a propensity that leads them to attend to rules linking events. Recent discoveries on mirror neurons (Rizzolatti 2005; Iacoboni, Molnar-Szakacs, Gallese, Buccino, Mazziotta & Rizzolatti 2005; Rizzolatti & Craighero 2004; Rizzolatti, Fogassi & Gallese 2000), for instance, indicate that primates do perceive meaningful actions performed by others by referring to the motor activation that would be generated if they were to perform the same action that they are perceiving and understanding. Activation of mirror neurons seems to be a requirement for understanding action (Falck-

Ytter, Gredebäck & Von Hofsten 2006) and may also be relevant for the perception of speech (Liberman & Mattingly 1985; Liberman, Cooper, Shankweiler & Studdert-Kennedy 1967).

The general importance of implicit information and our ability to extract embedded rules from experience with recurrent patterns has been recently demonstrated in experiments exploring more or less complex computer games with hidden principles (Holt & Wade 2005). The situation created by these games is structurally similar to the typical adult-infant interaction situations described above, with the addition that in the actual adult-infant interaction, the adult tends to help the infant by responding on-line to the infant's focus of attention and making the "hidden" rules much more explicit than in a game-like situation. At the same time, the complexity of the ecological scenario of adult-infant interaction is potentially much larger than what often is achieved in a computer game situation. But in this respect the infant's own perceptual and production limitations will actually help constraining the vast search space that it has to deal with. This is why some of the experimental evidence on infant speech perception, speech production and on the interaction between adults and infants was reviewed here. The combined message from these three areas of research all converge towards an image of the infant as being exposed to and processing its ambient language in a differentiated fashion. The recurrent asymmetry favouring high-low distinctions in infant perception and production, along with the adult's interpretation and reinforcement of that asymmetry, provides a relevant bias for the infant's future language development as the acoustic signal may effectively be structured along this dominant dimension already in the early stages of the language acquisition process. However, going from exposure to speech signals to the acquisition of linguistic structure is not possible using acoustic information alone. An emerging linguistic structure must necessarily rely on the very essence of the speech communication process, i.e., the link between the acoustic nature of speech and other sensory dimensions. Since these aspects are necessarily interwoven in real-life data, a simplified model of the early stages of the language acquisition process was created to study the impact of simple assumptions. An important venue of further model work is to investigate the basic constraints enabling the emergence

of grammatical structure. By systematic control of the model constraints and its linguistic input, it may be possible to get some insight on the necessary and sufficient conditions for higher levels of linguistic development. At any rate, future mismatches between the model predictions and empirical data will hopefully lead to a better understanding of the real language acquisition process.

References

Bechara, A.; Damasio, H.; Tranel, D.; Damasio, A. R. 1997. Deciding Advantageously Before Knowing the Advantageous Strategy. *Science.* **275**: 1293-1295.

Bechara, A.; Damasio, H.; Tranel, D.; Damasio, A. R. 2005. The Iowa Gambling Task and the somatic marker hypothesis: some questions and answers. *Trends in Cognitive Sciences.* **9:** 159-162.

Bertoncini, J.; Bijeljac-Babic, R.; Blumstein, S. E.; Mehler, J. 1987. Discrimination in neonates of very short CVs. *Journal of the Acoustical Society of America.* **82:** 31-37.

Best, C. T.; McRoberts, G. W.; Goodell, E. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America.* **109:** 775-794.

Bickerton, D. 1990. *Language and species*. Chicago: The University of Chicago Press.

Bruce, G. 1982. Textual aspects of prosody in Swedish. *Phonetica.* **39:** 274-287.

Cheney, D. L.; Seyfarth, R. M. 1990. *How monkeys see the world inside the mind of another species*. Chicago: The University of Chicago Press.

Daeschler, E. B.; Shubin, N. H.; Jenkins, F. A. 2006. A Devonian tetrapod-like fish and the evolution of the tetrapod body plan. *Nature.* **440:** 757-763.

Davis, B. L.; Lindblom, B. 2001. Phonetic Variability in Baby Talk and Development of Vowel Categories. In: F. Lacerda; C. von Hofsten; M. Heimann (Eds.). *Emerging Cognitive Abilities in Early Infancy*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 135-171.

Davis, B. L.; MacNeilage, P. 1995. The articulatory basis of babbling. *Journal of Speech and Hearing Research.* **38:** 1199-1211.

Dawkins, R. 1987. *The Blind Watchmaker*. New York: W.W. Norton & Company.

Dawkins, R. 1998. *Unweaving the Rainbow*. London: The Penguin Group.

Deacon, T. W. 1997. *The symbolic species: the co-evolution of language and the brain.* New York: W.W. Norton.

Degonda, N.; Mondadori, C. R.; Bosshardt, S.; Schmidt, C. F.; Boesiger, P.; Nitsch, R. M.; Hock, C.; Henke, K. 2005. Implicit associative learning engages the hippocampus and interacts with explicit associative learning. *Neuron.* **46:** 505-520.

Eimas, P. D. 1974. Auditory and linguistic processing of cues for place of articulation by infants. *Perception and Psychophysics*. **16:** 513-521.

Eimas, P. D.; Siqueland, E. R.; Jusczyk, P.; Vigorito, J. 1971. Speech perception in infants. *Science*. **171:** 303-306.

Enquist, M.; Arak, A.; Ghirlanda, S.; Wachtmeister, C. A. 2002. Spectacular phenomena and limits to rationality in genetic and cultural evolution. *Philosophical Transactions of The Royal Society B: Biological Sciences*. **357:** 1585-1594.

Enquist, M.; Ghirlanda, S. 1998. Evolutionary biology. The secrets of faces. *Nature*. **394:** 826-827.

Enquist, M.; Ghirlanda, S. 2005. *Neural Networks and Animal Behavior*. Princeton/ Oxford: Princeton University Press.

Falck-Ytter, T.; Gredebäck, G.; Von Hofsten, C. 2006. Infants predict other people's action goals. *Nature Neuroscience*. **9:** 878-879.

Fant, G. 1960. *Acoustic theory of speech production*. The Hague: Mouton.

Fernald, A.; Taeschner, T.; Dunn, J.; Papousek, M.; de Boysson-Bardies, B.; Fukui, I. 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*. **16:** 477-501.

Fort, A.; Manfredi, C. 1998. Acoustic analysis of newborn infant cry signals. *Medical Engineering & Physics*. **20:** 432-442.

Futuyama, D. J. 1998. *Evolutionary Biology*. Sunderland, Massachusetts: Sinauer Associates, Inc.

Gärdenfors, P. 2000. *Hur Homo blev sapiens: om tänkandets evolution*. Nora: Bokförlaget Nya Doxa.

Gay, T.; Lindblom, B.; Lubker, J. 1980. The production of bite block vowels: Acoustic equivalence by selective compensation. *Journal of the Acoustical Society of America*. **68:** S31.

Gay, T.; Lindblom, B.; Lubker, J. 1981. Production of bite-block vowels: Acoustic equivalence by selective compensation. *Journal of the Acoustical Society of America*. **69:** 802-810.

Gibson, E. J.; Pick, A. D. 2000. *An Ecological Approach to Perceptual Learning and Development*. Oxford: Oxford University Press.

Gil-da-Costa, R.; Palleroni, A.; Hauser, M. D.; Touchton, J.; Kelley, J. P. 2003. Rapid acquisition of an alarm response by a neotropical primate to a newly introduced avian predator. *Proceedings of The Royal Society B: Biological Sciences*. **270:** 605-610.

Gould, S. J. 1977. *Ontogeny and phylogeny*. Cambridge, Massachusetts: Belknap Press of Harvard University Press.

Hauser, M. D. 1989. Ontogenetic changes in the comprehension and production of Vervet monkey (Cercopithecus aethiops) vocalizations. *Journal of Comparative Psychology*. **103:** 149-158.

Hauser, M. D. 1992. Articulatory and social factors influence the acoustic structure of rhesus monkey vocalizations: a learned mode of production? *Journal of the Acoustical Society of America*. **91:** 2175-2179.

Hauser, M. D.; Chomsky, N.; Fitch, W. T. 2002. The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*. **298:** 1569-1579.

Hauser, M. D.; Wrangham, R. W. 1987. Manipulation of food calls in captive chimpanzees. A preliminary report. *Folia Primatologica*. **48:** 207-210.

Holt, L. L.; Wade, T. 2005. Categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *Journal of the Acoustical Society of America*. **117(4, Pt2):** 2621.

Houston, D.; Jusczyk, P.; Kuijpers, C.; Coolen, R.; Cutler, A. 2000. Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin and Review*. **7:** 504-509.

Iacoboni, M.; Molnar-Szakacs, I.; Gallese, V.; Buccino, G.; Mazziotta, J. C.; Rizzolatti, G. 2005. Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology*. **3(3):** e79.

Iverson, P.; Kuhl, P. 1995. Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*. **97:** 553-562.

Iverson, P.; Kuhl, P. 2000. Perceptual magnet and phoneme boundary effects in speech perception: do they arise from a common mechanism? *Perception & Psychophysics*. **62:** 874-886.

Jacob, F. 1982. *The possible and the actual*. New York: Pantheon.

Jakobson, R. 1968. *Child language. Aphasia and phonological universals*. The Hague: Mouton [72 ed.].

Johansson, S. 2005. *Origins of language : constraints on hypotheses*. Amsterdam/ Philadelphia, PA: John Benjamins.

Johnson, E. K.; Jusczyk, P. 2001. Word segmentation by 8-month-olds: when speech cues count more than statistics. *Journal of Memory and Language*. **44:** 548-567.

Jones, S.; Martin, R.; Pilbeam, D. 1992. *The Cambridge Encyclopedia of Human Evolution*. Cambridge: Cambridge University Press.

Jusczyk, P. 1999. How infants begin to extract words from speech. *Trends in Cognitive Sciences*. **3:** 323-328.

Kamo, M.; Ghirlanda, S.; Enquist, M. 2002. The evolution of signal form: effects of learned versus inherited recognition. *Proceedings of The Royal Society B: Biological Sciences*. **269:** 1765-1771.

Kelly, G. A. 1963. *A theroy of personality: The psychology of personal constructs*. New York: W. W. Norton.

Knight, C.; Studdert-Kennedy, M.; Hurford, J. R. 2000. Language: A Darwinian adaptation? In: C.Knight (Ed.). *Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*. New York: Cambridge University Press, 1-15.

Kuhl, P. K. 1985. Categorization of speech by infants. In: J. Mehler; R. Fox (Eds.). *Neonate Cognition: Beyond the Blooming Buzzing Confusion*. Hillsdale, New Jersey: Laurence Earlbaum Associates, 231-262.

Kuhl, P. 1991. Human adults and human infants show a perceptual magnet effect

for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*. **50:** 93-107.

Kuhl, P.; Andruski, J. E.; Chistovich, I. A.; Chistovich, L. A.; Kozhevnikova, E. V.; Ryskina, V. L.; Stolyarova, E. I.; Sundberg, U.; Lacerda, F. 1997. Cross-language analysis of phonetic units in language addressed to infants. *Science*. **277:** 684-686.

Kuhl, P.; Meltzoff, A. N. 1982. The bimodal perception of speech in infancy. *Science*. **218:** 1138-1141.

Kuhl, P.; Meltzoff, A. N. 1984. The intermodal representation of speech in infants. *Infant Behavior and Development*. **7:** 361-381.

Kuhl, P.; Meltzoff, A. N. 1988. The bimodal perception of speech in infancy. *Science*. **218:** 1138-1141.

Kuhl, P.; Meltzoff, A. N. 1996. Infant vocalizations in response to speech: vocal imitation and developmental change. *Journal of the Acoustical Society of America*. **100:** 2425-2438.

Kuhl, P.; Williams, K.; Lacerda, F.; Stevens, K. N.; Lindblom, B. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*. **255:** 606-608.

Lacerda, F. 1992a. Does Young Infants Vowel Perception Favor High-Low Contrasts? *International Journal of Psychology*. **27:** 58-59.

Lacerda, F. 1992b. Young Infants' Discrimination of Confusable Speech Signals. In: M. E. H.Schouten (Ed.). *The Auditory Processing of Speech: From Sounds to Words*. Berlin: Mouton de Gruyter, 229-238.

Lacerda, F. 2003. Phonology: An emergent consequence of memory constraints and sensory input. *Reading and Writing: An Interdisciplinary Journal*. **16**: 41-59.

Lacerda, F.; Ichijima, T. 1995. Adult judgements of infant vocalizations. In: K. Ellenius; P. Branderud (Eds.). *Proceedings of the XIIIth International Congress of Phonetic Sciences*. Stockholm: KTH/Stockholm University, vol. 1, 142-145.

Lacerda, F.; Klintfors, E.; Gustavsson, L.; Lagerkvist, L.; Marklund, E.; Sundberg, U. 2004. Ecological Theory of Language Acquisition. In: L. Berthouze; H. Kozima; C. G. Prince; G. Sandini; G. Stojanov; C. Balkenius (Eds.). *Proceedings of the Forth International Workshop on Epigenetic Robotics*. Lund University Cognitive Studies. **117:** 147-148.

Lacerda, F.; Marklund, E.; Lagerkvist, L.; Gustavsson, L.; Klintfors, E.; Sundberg, U. 2004. On the linguistic implications of context-bound adult-infant interactions. In: L. Berthouze; H. Kozima; C. G. Prince; G. Sandini; G. Stojanov; C. Balkenius (Eds.). *Proceedings of the Forth International Workshop on Epigenetic Robotics*. Lund University Cognitive Studies. **117:** 149-150.

Lacerda, F.; Sundberg, U. 1996. Linguistic strategies in the first 6-months of life. *Journal of the Acoustical Society of America*. **100:** 2574.

Lacerda, F.; Sundberg, U.; Klintfors, E.; Gustavsson, L. forthcoming. Infants learn nouns from audio-visual contingencies. Unpublished.

Lane, H.; Denny, M.; Guenther, F. H.; Matthies, M. L.; Menard, L.; Perkell, J. S.; Stockmann, E.; Tiede, M.; Vic, J.; Zandipour, M. 2005. Effects of bite blocks and hearing status on vowel production. *Journal of the Acoustical Society of America*. **118:** 1636-1646.

Liberman, A.; Cooper, F. S.; Shankweiler, D. P.; Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychological Review.* **74:** 431-461.

Liberman, A.; Mattingly, I. 1985. The motor theory of speech perception revised. *Cognition.* **21:** 1-36.

Liljencrants, J.; Lindblom, B. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language.* **48:** 839-862.

Lindblom, B.; Lubker, J.; Gay, T. 1977. Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of the Acoustical Society of America.* **62:** S15.

Locke, J. L. 1996. Why do infants begin to talk? Language as an unintended consequence. *Journal of Child Language.* **23:** 251-268.

Lotto, A. J.; Kluender, K. R.; Holt, L. L. 1998. Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America.* **103:** 3648-3655.

MacNeilage, P.; Davis, B. L. 2000a. Deriving speech from nonspeech: a view from ontogeny. *Phonetica.* **57(2-4):** 284-296.

MacNeilage, P.; Davis, B. L. 2000b. On the Origin of Internal Structure of Word Forms. *Science.* **288:** 527-531.

Maddieson, I. 1980. Phonological generalizations from the UCLA Phonological Segment Inventory Database. *UCLA Working Papers in Phonetics.* **50:** 57-68.

Maddieson, I.; Emmorey, K. 1985. Relationship between semivowels and vowels: cross-linguistic investigations of acoustic difference and coarticulation. *Phonetica.* **42:** 163-174.

Maia, T. V.; McClelland, J. L. 2004. A reexamination of the evidence for the somatic marker hypothesis: what participants really know in the Iowa gambling task. *Proceedings of the National Academy of Sciences of the United States of America.* **101:** 16075-16080.

Mandel, D.; Jusczyk, P. 1996. When do infants respond to their names in sentences? *Infant Behavior and Development.* **19:** 598.

Mandel, D.; Kemler Nelson, D. G.; Jusczyk, P. 1996. Infants remember the order of words in a spoken sentence. *Cognitive Development.* **11:** 181-196.

Mattys, S. L.; Jusczyk, P. 2001. Do infants segment words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance.* **27:** 644-655.

Meltzoff, A. N.; Borton, R. W. 1979. Intermodal matching by human neo-nates. *Nature.* **282:** 403-404.

Meltzoff, A. N.; Moore, M. K. 1977. Imitation of facial and manual gestures by human neonates. *Science.* **198:** 75-78.

Meltzoff, A. N.; Moore, M. K. 1983. Newborn infants imitate adult facial gestures. *Child Development.* **54:** 702-709.

Menard, L.; Schwartz, J. L.; Boe, L. J. 2004. Role of vocal tract morphology in speech development: perceptual targets and sensorimotor maps for synthesized French vowels from birth to adulthood. *Journal of Speech, Language, and Hearing Research.* **47:** 1059-1080.

Munakata, Y.; Pfaffly, J. 2004. Hebbian learning and development. *Developmental Science.* **7:** 141-148.

Nazzi, T.; Jusczyk, P.; Johnson, E. K. 2000. Language Discrimination by English-Learning 5-Month-Olds: Effects of Rhythm and Familiarity. *Journal of Memory and Language.* **43:** 1-19.

Nishimura, T.; Mikami, A.; Suzuki, J.; Matsuzawa, T. 2006. Descent of the hyoid in chimpanzees: evolution of face flattening and speech. *Journal of Human Evolution.* **51(3):** 244-254.

Nowak, M. A.; Komarova, N.; Niyogi, P. 2001. Evolution of Universal Grammar. *Science.* **291:** 114-118.

Nowak, M. A.; Plotkin, J. B.; Jansen, V. A. A. 2000. The evolution of syntactic communication. *Nature.* **404:** 495-498.

Nylin, S. (6-7-2006). Personal Communication.

Papousek, M.; Hwang, S. F. C. 1991. Tone and intonation in Mandarin babytalk topresyllabic infants: Comparison with registers of adult conversation and foreign languageinstruction. *Applied Psycholinguistics.* **12:** 481-504.

Papousek, M.; Papousek, H. 1989. Forms and functions of vocal matching in interactions between mothers and their precanonical infants. *First Language.* **9:**137-158.

Polka, L.; Bohn, O. S. 2003. Asymmetries in vowel perception. *Speech Communication.* **41:** 221-231.

Polka, L.; Werker, J. F. 1994. Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance.* **20:** 421-435.

Rizzolatti, G. 2005. The mirror neuron system and its function in humans. *Anatomy and Embryology.* **210:** 419-421.

Rizzolatti, G.; Craighero, L. 2004. The mirror-neuron system. *Annual Review of Neuroscience.* **27:** 169-192.

Rizzolatti, G.; Fogassi, L.; Gallese, V. 2000. Mirror neurons: Intentionality detectors? *International Journal of Psychology.* **35:** 205.

Robb, M. P.; Cacace, A. T. 1995. Estimation of formant frequencies in infant cry. *International Journal of Pediatric Otorhinolaryngology.* **32:** 57-67.

Roug, L.; Landberg, I.; Lundberg, L. J. 1989. Phonetic development in early infancy: a study of four Swedish children during the first eighteen months of life. *Journal of Child Language.* **16:** 19-40.

Saffran, J. R. 2002. Constraints on statistical language learning. *Journal of Memory and Language.* **47:** 172-196.

Saffran, J. R. 2003. Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science.* **12:** 110-114.

Saffran, J. R.; Aslin, R. N.; Newport, E. 1996. Statistical learning by 8-month old infants. *Science.* **274:** 1926-1928.

Saffran, J. R.; Thiessen, E. D. 2003. Pattern induction by infant language learners. *Developmental Psychology.* **39:** 484-494.

Savage-Rumbaugh, E. S.; Murphy, J.; Sevcik, R. A.; Brakke, K. E.; Williams, S. L.; Rumbaugh, D. M. 1993. Language comprehension in ape and child. *Monographs of the Society for Research in Child Development.* **58:** 1-222.

Schick, B. S.; Marschark, M.; Spencer, P. E.; Ebrary, I. 2006. *Advances in the sign language development of deaf children*. Oxford: Oxford University Press.

Seidenberg, M. S.; MacDonald, M. C.; Saffran, J. R. 2002. Does grammar start where statistics stop? *Science.* **298:** 553-554.

Stern, D. N.; Spieker, S.; Barnett, R. K.; MacKain, K. 1983. The prosody of maternal speech: Infant age and context related changes. *Journal of Child Language.* **10:** 1-15.

Stevens, K. N. 1998. *Acoustic Phonetics*. Cambridge, Mass.: MIT Press.

Sundberg, U. 1998. *Mother tongue - Phonetic Aspects of Infant-Directed Speech.* Stockholm: Stockholm University.

Sundberg, U.; Lacerda, F. 1999. Voice onset time in speech to infants and adults. *Phonetica.* **56:** 186-199.

Tees, R. C.; Werker, J. F. 1984. Perceptual flexibility: maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology.* **38:** 579-590.

Tomasello, M.; Carpenter, M. 2005. The emergence of social cognition in three young chimpanzees. *Monographs of the Society for Research in Child Development.* **70:** vii-132.

Tomasello, M.; Savage-Rumbaugh, S.; Kruger, A. C. 1993. Imitative learning of actions on objects by children, chimpanzees, and enculturated chimpanzees. *Child Development.* **64:** 1688-1705.

Van der Weijer, J. 1999. *Language Input for Word Discovery.* MPI Series in Psycholinguistics.

Werker, J. F.; Gilbert, J. H. V.; Humphrey, K.; Tees, R. C. 1981. Developmental aspects of cross-language speech perception. *Child Development.* **52:** 349-355.

Werker, J. F.; Logan, J. S. 1985. Cross-language evidence for three factors in speech perception. *Perception and Psychophysics.* **37:** 35-44.

Werker, J. F.; Tees, R. C. 1983. Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology.* **37:** 278-296.

Werker, J. F.; Tees, R. C. 1992. The organization and reorganization of human speech perception. *Annual Review of Neuroscience.* **15:** 377-402.

Werner, L. 1992. Interpreting Developmental Psychoacoustics. In: L.Werner; E. Rubel (Eds.). *Develpmental Psycholinguistics* Washington: American Psychological Association, 47-88 [1 ed.].

Werner, S. (3-15-2006). Cellular density in human tissues. Personal Communication.