Emergence of words

Multisensory precursors of sound-meaning associations in infancy

Eeva Klintfors
To Betty

I hope that you will one day find a profession that gives you the same happiness and satisfaction as mine does.
Abstract


This thesis presents four experimental studies, carried out at the Phonetic laboratory, Stockholm University, on infants’ ability to establish auditory-visual sound-meaning associations as a precursor of early word acquisition. Study I reports on the effect of linguistic variance on infants’ ability (3- to 20-months) to establish sound-meaning associations. The target-words embedded in phrases, based on an artificial language, were presented along with visually displayed puppets. Study II investigates the role of attribute type on infants’ ability (3- to 6-months) to establish sound-meaning associations. Two-word phrases, based on the same artificial language as in Study I, were presented along with visually displayed geometrical objects. The words implicitly referred to the color and shape of the objects. Study III examines infants’ ability (12- to 16-months) to predict phonetic information. The subjects were tested on their ability to associate Swedish whole words and disrupted words to familiar objects. Study IV investigates infants’ ability (6- to 8-months) to detect concurrence and synchrony in speech and non-speech. The infants were exposed to Swedish speech sounds presented with corresponding articulatory events, the sound of hand-clapping presented with synchronized hand-clapping movements, and the sound of hand-clapping presented with synchronized articulatory events. The results picture the subject as sensitive to distributional properties of auditory and visual information (Study I and II) but still unable to predict phonetic information, in the beginning of the second year of life (Study III). The infants’ conceptual behavior is outlined as a general-purpose perceptual process influenced by perceptual salience (Study IV). These results are related to a working hypothesis based on the Ecological theory of language acquisition (Lacerda & Sundberg, 2006), and Lindblom (Lindblom, 1990; Lindblom & Lacerda, 2006).

Keywords: language acquisition; word acquisition; auditory-visual; multisensory; sound-meaning association; distributional learning; perceptual salience; infancy
List of studies


Data collection was carried out by E. Klintfors and L. Gustavsson (Study I), E. Klintfors and E. Marklund (Study II), and students in logopedics supervised by F. Lacerda and E. Klintfors (Study III, IV). The statistical analysis was performed by E. Klintfors and F. Lacerda (Study I-IV). The preparation of the manuscripts was performed by E. Klintfors (Study I, II, III) and F. Lacerda and E. Klintfors (Study IV). Study IV consists of two sub-studies. The sub-study ‘Linking audio-visual information’ is one of the present dissertation studies, while the sub-study ‘Emerging word-object associations’ was mainly carried out by L. Gustavsson and F. Lacerda and is therefore not addressed in this thesis.
Other publications by the author

Koponen, E., Kasaty, A. (1998): Swedish nominal morphophonology implemented within the two-level model in PC-Kimmo, Bachelor of Arts thesis in computational linguistics, Department of linguistics, SU.

Koponen, E. (1999): Effects of target-word frequency rate on sound-meaning-connection in five to fifteen month-old Swedish infants, Master of Arts thesis in computational linguistics/Bachelor of Arts thesis in phonetics, Department of linguistics, SU.


Koponen, E. (2001) Assessing the relevance of prosodic and phonotactic cues on parsing the speech stream by young language-learners, Master of Arts thesis in phonetics, Department of linguistics, SU.


Gustavsson, L., Sundberg, U., Klintfors, E., Marklund, E., Lagerkvist, L. and Lacerda, F. (2004) Integration of audio-visual information in 8-months-


Acknowledgements

This work was supported by ERICSSON and Faculty of Humanities through the project Späd barnets Språkvärld, by Swedish Research Council through the project Språkets Begynnelse, by Bank of Sweden Tercentenary Foundation through the project Modeling Interactive Language Learning (MILLE), by European Commission through the project Learning and Development of Contextual Action (CONTACT), and by Birgit and Gad Rausing’s Foundation through the project Vad Lyssnar Späd barn på När vi Talar till dem?. Elisabeth and Herman Rhodin’s Foundation, Anna Ahlström and Ellen Terserus Foundation, Department of Linguistics, Stockholm University (SU) and European Commission have contributed by part time doctoral student scholarships and employments for doctoral studies. Board of Department of Linguistics, SU, Wallenberg Foundation, International Speech Communication Association (ISCA) and Einar and Carolina Bergström Foundation have supplied with several traveling contributions.

First of all I would humbly like to thank my supervisor, the principle investigator of the current research projects at the Babylab, SU, Francisco Lacerda. His long experience within engineering, phonetics, theory of science, statistics and especially language development was invaluable for the quality of this thesis and the experiments performed. Thank you Frasse for the one-of-a-kind enthusiasm you spread around, for all the devoted discussions at the University and other places around the world, and for being such a generous person. You are the best supervisor and coach into the academic life I can think of!

I would also like to thank the other of my two closest colleagues, the person who has been my unofficial mentor right from the start, Ulla Sundberg. Ulla planned and supervised all the work together with Frasse. Her all-round skills in child development in general, and language development, speech therapy, and infant-directed speech in particular, were crucial for the whole Babylab-team to keep track of the latest literature and experiments done within the field. Thank you Ulla for reading and commenting on a number of papers I have written during the last years, for being such a fun traveling company, and for always being there to talk about life, work and carriers!
Going back to the beginning, thank you especially Björn Lindblom for functioning as an informal supervisor already from the basic level course in linguistics, for the influential C-level course in phonetics you gave your last teaching term at SU, and for recommending me as an assistant teacher. I am also grateful to (late) Gunnel Källgren for sending me to Bertil Lyberg, Anders Lindström, Robert Eklund and others at Telia Research to get the best available introduction to the field of speech technology, for suggesting to combine computational linguistics with phonetics, and for encouraging me to continue to graduate school. Thank you (again) Frasse, for captivatingly introducing me to the literature of emergent phonology, for presenting research problems for which investigation computational skills were of benefit, and for offering me a research assistant position. Thank you also Benny Brodda for supporting the idea of using computational methods to study phonetic research questions, Ulla (again) for many insightful initial discussions of emergent phonology, and Frasse, Ulla, and Olle Engstrand for not giving up in search for, and in supporting my efforts when applying for financing to carry out doctoral studies. 

I am also grateful for the opportunity I was given to obtain teaching and administrative experience at the department of Linguistics. Thank you Christina Hellman and Päivi Juvonen for putting your faith in me in letting me teach and develop materials for the basic and advanced levels in phonetics. In addition I would like to express my appreciation to my former teachers in particular (Björn Lindblom, Olle Engstrand, Masja Koptjevskaja Tamm, Mikael Parkvall, Östen Dahl, Christina Hellman, Hartmut Traummüller, Bob McAllister, Frasse, (late) Gunnel Källgren, Benny Brodda, Sofia Gustafson-Capková, Jennifer Spenader, Beá(ta) Megyesi, Sara Rydin, Harald Berthelsen, etc.) and to colleagues and co-doctoral students at the department in general. A special ‘kiitos’ goes to my compatriot friend and former neighbor Päivi for close cooperation and continuous advising in tutoring individual students with diverse needs. The administrative staff (Linda Habermann, Lotta Vouletthe, Cilla Nilsson, Nada Djokic, Liisa Karhapää), as well as the technician staff (Bosse Kassling, Hassan Djamshidpey, Peter Branderud) of the department deserve a hug for their friendliness and helpfulness in solving everyday problems.

Within MILLE and CONTACT I had the great opportunity to meet and work together with several highly skilled researchers from around the world. The MILLE-project, which is an interdisciplinary research collaboration between three research groups – Department of Linguistics (SU, Sweden), Department of Psychology, Carnegie Mellon University (CMU, USA) and Department of Speech, Music and Hearing, Royal Institute of Technology (KTH, Sweden) – has led to close cooperation with Lori Holt, Christi Gomez, and (An)thony Kelly from CMU, as well as with Rolf Carlson, Björn
Granström, Kjell Elenius, and Giampiero Salvi from KTH. Thank you especially Lori (with family) for taking such a good care of us during visits to Pittsburgh! Within CONTACT, which is a European co-work between – Department of Psychology, Uppsala University (UU, Sweden), Department of Linguistics (SU, Sweden), Department of Communication, Computer and System Sciences/Laboratory for Integrated Advanced Robotics, University of Genova (UNIGE, Italy), Neurolab/Department of Biomedical Sciences-Human Physiology, University of Ferrara (UNIFE, Italy) and VisLab/Department of Electrical and Computer Engineering, Instituto Superior Técnico of Lisbon (IST, Portugal) – I had the pleasure to collaborate with the following people among others: Claes von Hofsten, Kerstin Rosander, Ben(jamin) Kenward, Hélena Grönqvist, Pär Nyström, Helena Örnkloo, Olga Kochukhova, Giulio Sandini, Paul Fitzpatrick, Giorgio Metta, Laila Craighero, Luciano Fadiga, José Santos-Victor, Manuel Lopes, and Jonas Hörnstein.

In addition to external collaborators, the whole Babylab-team (including permanent staff and occasional visitors) has provided the most inspiring environment for research. I also have several people to thank for their help in reading and improving parts of this thesis, especially Ulla Sundberg, Ljuba Veselinova, Östen Dahl, Mikael Parkvall, Anders Eriksson, Iris-Corinna Schwarz, Brita Bergman, Björn Lindblom, Ben Kenward and Ellen Marklund (who prepared some figures and carried out final formatting).

Further importantly, I wish to thank the infant-subjects and their parents for participation in the experiments, for patience with initial methodological plights, and for interest shown in the research questions studied.

Finally, I send my love to my dear family – Rolf and Betty Klintfors. Thank you for providing peaceful surroundings conducive to work. Your patience with and support to ‘mamma’ has been highly important. You make my life special!
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETLA</td>
<td>Ecological theory of language acquisition</td>
</tr>
<tr>
<td>S-R</td>
<td>Stimulus-response</td>
</tr>
<tr>
<td>UG</td>
<td>Universal grammar</td>
</tr>
<tr>
<td>ASL</td>
<td>American sign language</td>
</tr>
<tr>
<td>IDS</td>
<td>Infant-directed speech</td>
</tr>
<tr>
<td>SLI</td>
<td>Specific language impairment</td>
</tr>
<tr>
<td>ASD</td>
<td>Autism spectrum disorder(s)</td>
</tr>
<tr>
<td>FSG</td>
<td>Final state grammar</td>
</tr>
<tr>
<td>PSG</td>
<td>Phrase structure grammar</td>
</tr>
<tr>
<td>H&amp;H</td>
<td>Hyper &amp; Hypo (theory)</td>
</tr>
</tbody>
</table>
Part I Introduction
Chapter 1

Aims and outline

The primary goal of this thesis is to focus on fundamental processes involved in early language acquisition considering the multisensory character of infants’ interaction with their ecological environment. In particular, the present dissertation studies investigate how auditory-visual sound-meaning associations may function as a precursor of early word acquisition.

Our search for knowledge about the fundamental processes involved in acquiring the ambient language may be facilitated by regarding language acquisition in a broad interdisciplinary perspective. Within this framework, previous research representing various theoretical views is discussed in Part I Introduction (Chapter 2 Early word acquisition) by addressing the following topics: (1) What characterizes speech perception and conceptual behavior in infants vs. adults?, (2) How does speech perception relate to the perception of non-speech?, (3) How does typically developing infants’ speech perception differ from speech perception in atypically (e.g. hearing or visually impaired) developing infants?, and (4) What perceptual and cognitive skills do humans have that are useful for language, and to what extent are they shared by non-human animals? The aim of Part I Introduction is thus to discuss what is known about language development as well as to raise questions for further study. The author’s theoretical assumptions, based on ETLA (Lacerda & Sundberg, 2006) and Lindblom’s work (Lindblom, 1990; Lindblom & Lacerda, 2006) are presented here.

The four studies which constitute this thesis were primarily carried out within two parallel projects. The aim of the Bank of Sweden Tercentenary Foundation project MILLE was to investigate fundamental processes involved in language acquisition and to develop computational models to simulate these processes which are potentially similar for humans and other mammalian species. The European Commission project CONTACT aimed at building an embodied system (a humanoid baby robot) which develops its motor capabilities both in manipulative gestures and speech. A central hypothesis under investigation in CONTACT is that there are commonalities between the development of speech and manipulation. The theoretical starting point of MILLE and CONTACT, based on ETLA (Lacerda & Sundberg, 2006), was that emergence of initial linguistic referential function
may occur as the child interacts within her/his ecological environment and might be accounted for on the basis of general-purpose perceptual, production and memory processes.

The specific objectives of the dissertation studies are multifold. The aim of the first study is to explore the effect of linguistic variance on infants’ ability to establish auditory-visual sound-meaning associations. The speech materials were characterized by varying target-word frequencies and phrase positions. The goal of the second study is to investigate the role of attribute type on infants’ ability to establish auditory-visual sound-meaning associations. The attributes tested were shape and color of objects. The aim of the third study is to assess infants’ ability to derive sound-meaning associations from complete vs. incomplete phonetic information (intact vs. interrupted words) on a target-object, in an audio-visual naming task. The complete words referring to objects were presented in their entirety, while the incomplete words referring to objects were partially masked so that only initial phonetic information was presented. The aim of the forth study is to address infants’ ability to detect concurrence between synchronized speech and non-speech events. The materials were Swedish speech sounds presented with corresponding articulatory events, and the sound of hand-clapping presented with synchronized hand-clapping movements or articulatory events.

Knowledge of speech and language is fragmented across a wide range of disciplines such as phonetics, linguistics, psychology, philosophy, anthropology, cognitive science, neuroscience, computer science, artificial intelligence, engineering, robotics, etc. Integrating results from all these areas is in itself a major challenge for the whole research community. Therefore some aspects of language acquisition will only be briefly discussed. Also, the four parts of the thesis investigate early perceptual word acquisition from an ontogenetic perspective, and therefore issues dealing with word production,\(^1\) acquisition of grammar *per se*,\(^2\) and the evolutionary emergence of language fall outside its scope. In addition, interest in the underpinnings of intelligent human behavior within different disciplines comprises the study of mental representation,\(^3\) and of cognitive capacities as simulated by machines\(^4\) which are two other issues outside this thesis.

\(^1\) For example, emergence of first words, growth of vocabulary, articulatory aspects of word production, etc.
\(^2\) For example, phonology, morphology, case, agreement, syntactic recursion, etc.
\(^3\) For example, the way language is represented in the brain as looked into through neural-imaging.
\(^4\) For example, computational modeling of fundamental mechanisms underlying spoken language processing.
Part I Introduction (Chapter 2) discusses the following topics: 2.1 What does it mean to know a word?, 2.2 Interdisciplinary perspective on language acquisition, 2.3 Speech perception in infants vs. adults, 2.4 Perception of speech vs. non-speech, 2.5 Speech perception in typically vs. atypically developing infants, and 2.6 Speech perception in humans vs. non-human animals. These issues provide a selective overview of theoretical and empirical accounts of speech perception and language acquisition, followed by an evaluation of methodological and statistical aspects of the current research (Part II Contributions of the present work). Closing remarks and suggestions for future research conclude each chapter. The final chapter (Chapter 5 Discussion and conclusions) relates the results of the current research to the working hypothesis of this thesis. Finally, the present dissertation studies, published in peer-reviewed conference proceedings, are presented (Part III Studies).
Chapter 2

Early word acquisition

2.1 What does it mean to know a word?

This section introduces criteria for what is meant by knowing a word, which are needed for understanding discussions about the underlying mechanisms of word learning in this thesis. Partly, it is a review of related concepts such as cognition, consciousness, general/language knowledge, and categorization ability in relation to word learning. This section will also present the fundamentals of the working hypothesis on word meaning acquisition in this thesis. Categorization as a way of reducing the load on memory and other cognitive functions is exemplified. The last paragraph reviews what it means to know a word from the perspective of early speech production.

Knowledge may be defined as expertise or skills acquired through education or experience (Oxford English dictionary). Knowing a language comes about through experience and its user is often unaware of possessing that knowledge. Or to put it in Tulving’s words: “It is perfectly possible for people to engage in highly complicated forms of learned behaviour without anything resembling the corresponding knowledge. Young children’s ability to speak grammatically is one of the most obvious examples” (Tulving, 2000, p. 37).

Knowing what a word refers to, whether it is an object or an abstract phenomenon in the world, is also typically based on information received via experience. Even when the definition of a word is explicitly given in a dictionary or carefully explained by somebody, word learning involves creating mental categories, i.e. concepts of objects and events that have a common set of properties. These properties are usually based on sensory information. One may for example rely on physically varying continua like frequency and amplitude of electromagnetic waves in vision (conveying hue/color and intensity) or frequency and amplitude of pressure fluctuations in hearing (conveying pitch and loudness). But in contrast to receiving information through one sensory channel, learning the meaning of words in spoken language typically requires association of information from the auditory channel with correlated information from other sensory channels.
More explicitly, according to the working hypothesis of this thesis, word acquisition can be seen as part of a continuum, stretching from the integration of initially fuzzy sensory associations in newborns, to the establishment of sound-meaning associations by the age of about 6- to 12-months, and leading to the more sophisticated formation of abstract concepts at about 18- to 36-months. In this view, the newborn infant’s early postnatal conceptual behavior is assumed to build on initially fuzzy associations between multi-sensory (visual, auditory, olfactory, gustatory, tactile as well as kinaesthetic) information (Lacerda & Sundberg, 2006). Later on, by the age of about 6- to 12-months, the infant’s initially fuzzy sensory associations are assumed to converge onto a developmental stage portrayed as a rule-based capacity^5 to associate sound patterns to visually consistent objects. During this period the child’s word acquisition process is mainly characterized by labeling objects or actions within reach. Towards the end of the continuum, at about 18- to 36-months of age, the child’s existing vocabulary for objects and actions and her growing sensitivity to referential intentions of parents and other caregivers are assumed to play an increasingly important role in the emergence of words corresponding to more abstract concepts. Here a distinction is made between common nouns corresponding to things and objects which can, or cannot, be defined by their physical attributes. For example, consider the concept ‘food’. It is the use of a hamburger and a piece of chocolate, as opposed to physical similarity, that links them together. Thus, when an object is defined by its functions, the word clearly corresponds to a more abstract concept.

According to several associationist theories (Hull, 1920; Skinner, 1957; Harlow, 1959; Bourne & Restle, 1959; Gibson, 1940) the categorization process involves observing substantial properties that are characteristic for a certain category and distinguish it from other categories, while ignoring properties that are irrelevant for a certain category. For example, a red triangle is singled out from other objects particularly by its color and form,
while ignoring its size (Bourne, 1966). The ability to categorize is further assumed to reduce the load on memory and other cognitive processes (Rosch, 1975). A possible outcome of perceiving events and objects as unique is that our interaction with the environment might become complex and unpredictable (Bourne, 1966) (for a discussion of language learning disabilities in children presumably attributed to reduced or absent ability to generalize information across settings, see 2.5 Speech perception in typically vs. atypically developing infants).

As opposed to the associationist point of view, adherents of theories based on hypothesis testing (Krechevsky, 1932; Levine, 1959; Bruner et al., 1956) maintain that “First, he [the subject] may not respond to all available stimulus features but rather select and attend to only certain aspects which, on the basis of a hypothesis, are considered relevant. Second, the subject decides upon and executes a response, in conformance with the hypothesis, which serves as a test of its adequacy” (Bourne, 1966, p. 36). Associations between stimulus properties and responses based on hypotheses are further assumed to develop quickly (sometimes at one trial), and the essence of learning does not involve stimulus-response (S-R) associations but rather the recognition of a principle required by the task as in the case of rule-based learning.

In the current working hypothesis, the newborn may indeed focus on random, rather than on relevant stimulus properties, but assumptions made about irrelevant stimulus properties cease to exist since they are non-productive both in regards to further interactions with the infant’s immediate environment and to the infant’s own representations of her/his world. According to this, the infant’s ability to predict the environment increases with the development of conceptual behavior of the infant. Early assumption-testing can therefore be viewed as an unintended and implicit by-product of a general behavior involved in operating on the environment, rather than a goal-oriented enterprise. This is a suggestion that contrasts with Bruner and colleagues’ (1956) position that their subjects followed deliberate strategies towards conceptual behavior.

*Cognition,* which in mainstream psychology signifies mental processes dealing with thinking, memory, attention, perception, problem solving, *etc.*, is a term closely related to knowledge. Knowledge achieved through

---

7 For a discussion on what particular kind of irrelevant properties (*e.g.* allophonic variation caused by anatomy of the speaker) needs to be ignored to be able to categorize speech sound, see 2.5 Speech perception in typically vs. atypically developing infants. Also, for a discussion on how infants’ limited focus of attention may reduce the complexity of the ambient world, see Chapter 3 Distributional learning.

8 Latin *cogito*: I think.
experience \(e.g.\) via trial and error) often becomes specific to a particular situation, commonly referred to as \textit{situated knowledge}. This concept, among others deriving from the Gibsonian ecological approach to visual perception (Gibson & Pick, 2006), aims at describing how learning in the real world is always context-specific. Therefore, to fully explain a behavior, it is necessary to study the environment in which the addressed behavior takes place. In line with this assumption, the current learning theory presumes that infants’ observations of simultaneous word-object pairings may initially, just like other experience-based learning, lead to context-bound information and semantically wrong assumptions (Lacerda & Sundberg, 2006).

Awareness of knowledge is related to \textit{consciousness}\textsuperscript{9} which in common use simply designates being awake and responsive to one’s environment as opposed to being asleep or in coma. However in psychology, consciousness is often defined as the capacity for experience and the capacity for thoughts. The capacity for experience \(e.g.\) experience of pain is portrayed as a prerequisite of thought while conscious thoughts are taken to be more abstract and require the use of concepts (Giambrone & Povinelli, 2002). Following this terminology, many researchers investigating capacities related to consciousness among living taxa, agree that primates\textsuperscript{10} and several other non-human species might be aware of something in their environment or even be able to use concepts and think about them, but not necessarily to be conscious \(e.g.\) self-reflective of their own thoughts. In addition, \textit{reflective} self-consciousness has – among other cognitive capacities that separate humans from non-human primates – been suggested to be a prerequisite of language (Gärdenfors, 2000).

Finally, to answer the question of what it means to know a word is far from simple when one needs to define \textit{use} of earliest words. Studies of language learning in infants, struggling with the problem of word definition and identification, characteristically call upon two criteria: resemblance of phonetic form to an adult word form, and situational consistency in use (Lewis, 1936; Leopold, 1939; Nelson, 1973; Ferguson & Farwell, 1975; Greenfield & Smith, 1976). However, recognition of words in the earliest period of potential word production is further complicated by the fact that utterances by infants may be difficult to distinguish from concurrent babble (Vihman et al., 1985), and that meanings might be on the one hand unconventionally narrow\textsuperscript{11} and on the other hand global and diffuse.

\textsuperscript{9} Latin \textit{conscientia}: shared knowledge.
\textsuperscript{10} It has been suggested that the systematic connections underlying consciousness are present in primate brains only (Damasio, 1989).
\textsuperscript{11} For example, the word ‘car’ might be used to refer only to a single instance of the object in one particular context.
(Rescorla, 1980; Griffiths, 1986; Nelson, 1988). Also sound-meaning consistencies might initially be idiosyncratic (Halliday, 1975; Dore et al., 1976; Ferguson, 1978). The methodology for word identification suggested by Vihman and McCune (1994) uses criteria based on vocalization shape such as looking into the amount of segmental and/or prosodic consistency between child and adult forms. Criteria based on context in its turn focus on for example multiple use (does the child use the word more than once?) and multiple episodes (is there more than one episode of use?). In addition, Vihman and McCune’s (1994) formal system which evaluates phonetic match and functional categories reflecting use in context, relates a vocalization to other vocalizations by asking questions such as ‘Is there at least one instance imitated?’, or ‘Do all instances of the word exhibit the same phonological shape?’.

2.2 Interdisciplinary perspective on language acquisition

This section presents introductory material on research perspectives within natural and social sciences. These sciences contrast causal and teleological explanations. The present working hypothesis draws on both. Additionally, two well-known dichotomous views (nature vs. nurture) on language acquisition will be briefly introduced.

Strzygowski (1923) groups fields of science into two categories: factual-research and viewer-research – today commonly referred to as natural and social sciences. One of the main ideas that separate these, according to Strzygowski, is that natural sciences aim to explain whereas social sciences want to understand. Hence, explanations within natural sciences should not only tell us that a definite effect usually follows a specific cause – but help us to explain why the relation exists and give us the mechanism behind it. Such a mechanism is preferably accounted for with reference to independently motivated general biological, chemical, or physical phenomena. As opposed to causal explanations, teleological clarifications of what somebody does and why aim at understanding intentions of individuals. Especially within arts and literature, the researcher is welcomed not only to independently observe, but to give his own expert interpretation of the factual intention of the artist. Similarly, Kant (1994) views human beings as biological organisms or as goal-oriented creatures.

As stated in the aims and outline (Chapter 1), this thesis maintains that search for scientific knowledge about fundamental processes involved in language acquisition may be facilitated by regarding it in a broad inter-

---

12 For example, the word ‘mother’ might be used to refer to women in several contexts.
13 Greece telos = goal.
disciplinary research perspective. Language learning in infants is therefore discussed with reference to adults’ conceptual behavior, atypical or impaired language development, perception of sounds in general (as opposed to speech), and with reference to potential precursors of learning in non-human species.

More specifically, the working hypothesis outlined in this thesis suggests that language acquisition may be addressed as a particular case of a learning process capturing relations between different sensory dimensions. Thus, instead of focusing on the acoustic signal \textit{per se} (e.g. Saffran et al., 1996a) acquisition of spoken language is described here in terms of a general detection process of contingencies in the sensory representation space available to the infant. The interdisciplinary research issues addressed consequently highlight the infant’s prenatal exposure leading to an attunement towards speech and sounds, the plasticity of the brain suggesting a non-modular structure for language, the compensatory learning mechanisms in atypically developing infants that replace typical precursors of word learning, and discuss how potential perception-processing similarities between infants and non-human species (with reference to biological and physical facts) may shed light on critical aspects of early language acquisition. These assumptions on language acquisition comprise the methodological stance from which the companion empirical investigations follow. For sure, the nature of the building blocks of language acquisition is controversial. For example, it has been suggested that the organization of the brain is modular for language (Fodor, 1983), as well as that language acquisition is based on knowledge of certain preexisting representations of physics (Carey & Spelke, 1996; Spelke, 2000) (to be discussed in 2.4 Perception of speech \textit{vs.} non-speech). Indeed, children come to this world with many sophisticated abilities (see \textit{e.g.} the functional infrastructure for language by Bates, discussed in 2.4 Perception of speech \textit{vs.} non-speech) perhaps for learning special tasks, but theorizing about the nature of these issues is not possible until the bottom-up approach has been exhausted. Thus, in order to evaluate the modularist position or alleged innate language acquisition mechanisms, it is necessary to profoundly explore the impact of the multisensory experience.

Language acquisition in early infancy is obviously a complex process involving a number of parallel components and their intricate relationships. However, the current working hypothesis adopts a minimalist research strategy according to which the early word-learning process is investigated by exploring the potential explanatory bearing of rather parsimonious components. For example, a common criticism against ‘pure’ associationism is that statistical stability of sound-meaning associations may not always be apparent for the infant because adults do not constantly utter nouns
simultaneously as the child is attending to an object (e.g. the child might rather be looking at the parent). Therefore, children’s sensitivity to referential intentions of other people (theory of mind) has suggested to be the central mechanism by which word-learning may be accomplished (Bloom, 2000; Bloom, 2001; Markson & Bloom, 1997). Yet, rather crude associative processes may account for early concept formation because the world of the very young infant is radically different from the adult’s. From that perspective, the infant’s limited focus of attention (to be discussed in Chapter 3 Distributional learning) effectively reduces the complexity of the ambient world during the early stages of language acquisition. Certainly, according to the adult perspective, language builds on shared knowledge of the world and conventionalized expressions for existent or hypothesized objects and events, but this characterizes a much later phase in the development. Thus, as interpreted by ETLA, infants lack long-term experience of the world and language and might be happily unaware of the ‘goal’ of language learning. In addition to the suggested non-teleological learning perspective of the infant, the context-proximate early stages of infant-adult interactions, providing lots of systematic relations between acoustic labels and their referents (Lacerda & Sundberg, 2006), are considered to lead to a highly supportive learning environment for the infant.

Interpretation of language acquisition has for a long time been strongly influenced by two dichotomous views: language learning is regarded as either determined by an individual’s innate capacities (nature) or by personal experiences (nurture). Yet, several approaches today – based on evidence from theoretical physics, neural networks, and neurobiology – agree that genes and environment interact to determine complex cognitive outcomes. The working hypothesis outlined in this thesis acknowledges the importance of both genetic constraints and experience as explanatory factors in first language acquisition.

Learning theories can subsequently be classified, based on how organisms are presumed to operate on incoming information, along two main lines of

---

14 Also, children organize and reorganize their knowledge of the ambient language, frequently manifested by what is known as U-shaped learning (Strömqvist, 2008). For example, the child may first generalize a common plural ending (e.g. -ar) to a word stem (e.g. bok), then correctly produce the adult form (böck-er), followed by producing varying forms of the word stem (e.g. bok or böck), to finally settle with the correct stem (bok) and plural form (böcker) (Plunkett & Strömqvist, 1992). U-shaped learning has often led into considerations of children taking irrational detours in their language development. As pointed out by Strömqvist (2008), comparisons of children’s and adults’ language output overlook the fact that children are not aware of the contents of adult language competence. Instead, children act upon their (un)conscious experiences and factors in their immediate environment which guide the development of the child in a certain direction.
behavior (Heidbreder, 1924). According to the first, a learning organism is pictured as a passive recipient of information from the environment, whereas organism-internal activities make little or no contribution to concept formation. Proponents of behaviorism for example regard the acquisition of referential word function as a passive process wherein S-R, relying on external reinforcement, become gradually connected. As opposed to ‘pure’ associationist or behaviorist theories, the current working hypothesis regards the infant as a biological system, integrated in its environment. The infant is viewed as a participant who progressively becomes active, as exemplified by the indication that while infants younger than 16-months fail to use partial phonetic information, predictive behavior has been observed in infants in their second year of life (Fernald et al., 2001) (to be discussed in 4.1 Prediction of phonetic information). In addition, the current working hypothesis is an attempt to specify and understand the mechanisms behind infants’ assumption-testing about unknown concepts (or partial information). In summary, according to the terminology of Strzygowski and Kant, the interdisciplinary perspective of this thesis represents the non-teleological point of view within natural sciences, but also the standpoint of viewing human beings as acting organisms, as commonly adopted within social sciences.

Since the topic of this thesis is early word-learning and its undertaking is to place the companion studies of the establishment of sound-meaning associations in an interdisciplinary scientific context, I have chosen to organize the rest of this part (Part I Introduction) around issues of – Speech perception in infants vs. adults (2.3), Perception of speech vs. non-speech (2.4), Speech perception in typically vs. atypically developing infants (2.5) and Speech perception in humans vs. non-human animals (2.6) – rather than around common concerns (such as e.g. domain-specificity or species-specificity of language) within the nature-nurture-controversy.

### 2.3 Speech perception in infants vs. adults

This section begins with arguments – proposed by representatives of universal grammar (UG) – about why learning by imitation, analogy, or reinforcement cannot be sufficient mechanisms for language learning. The discussion further comprises experimental evidence (from a case study, from homesign systems, and from creole versions of pidgin languages) suggesting that children may indeed learn language from reduced or degraded input. This section goes then into more depth on what characterizes onset of conceptual behavior in children in particular. An important issue, according to the present working hypothesis, is the impact of prenatal learning for postnatal language acquisition. There is experimental evidence on prenatal olfactory, gustatory, and visual learning which may justify that the infant’s
postnatal recognition of and preference for speech might be a consequence of a general recognition process of auditory exposure in the uterus. Thus by arguing that infants’ rapid language development can partly be explained as a consequence of prenatal learning, this thesis offers an alternative account for the seemingly effortless language learning in typically developing infants.

Children acquire the remarkably complex system of language quickly and seemingly effortlessly. In typical development, explicit teaching of or specific motivation to learn the ambient language is hardly ever a pressing issue. These puzzling characteristics, in addition to the fact that most children demonstrate similar developmental stages of acquisition, have often led researchers to postulate an innate capacity for language. For example, Chomsky has proposed that “To come to know a human language would be an extraordinary intellectual achievement for a creature not specifically designed to accomplish this task” (1975, p. 4). Supporting Chomsky’s proposal of UG as an innate property of the human mind, children are often regarded as being unable to learn language through imitation, analogy, or reinforcement. Imitation is held as an insufficient way based on that children’s speech substantially differs from the speech spoken to them by adults. Examples of children not confining themselves to phrase structures that they have heard others say are for example provided by Pinker (1989). Analogy as a learning strategy, i.e. using some sentences as models for producing new ones, is in turn often ruled out based on that analogies are not applicable in all contexts. Finally, reinforcement by feedback is considered to be unlikely since adults seldom correct children when they speak ungrammatically (Pinker, 1989; Brown & Hanlon, 1970). Instead, adults seem to focus in general on the truth-value of the message, rather than on the grammatical correctness of the sentences uttered by children. The ‘Poverty of the stimulus’ or ‘Learnability paradox’ argument thus maintains that language is not learnable given the linguistic input available to the children. Children phrase new, correct sentences they never heard before, and they do so solely based on positive examples (i.e. examples of how to say correct phrases) but without explicit guidance to negative examples (i.e. examples of how not to say incorrect phrases). In line with these arguments demonstrating that the grammatical nature of speech directed to children is incomplete, Gold’s theorem (Gold, 1967) explicitly exemplifies that grammars with center embeddings and long-distance-dependencies cannot be learned in the absence of negative evidence. On the other hand, among critics of Gold’s theorem (Deacon, 1997; Hirsh-Pasek & Golinkoff, 1996; Cowie, 1997; Prinz, 2002), Elman (1993) has demonstrated that neural networks with children-like-constrained short-term

\[ \text{For a review of universal patterns in early language acquisition see Slobin (1973; 1985).} \]
memory may – after initial errors and later recovery from those errors – accomplish the task with help of a positive database only (i.e. without information on what possible sentences do not belong to the language in question).

Further evidence for children being able to inductively acquire the grammar of a language based on inadequate input data, comes from a case study of a deaf child called Simon (Singleton & Newport, 2004). Simon’s only American Sign Language (ASL) input was provided by his late-learner parents who, just like other late-learners of ASL, performed below adult native signing criteria. 16 Since no history of deafness was known in either of the parent’s families, Simon was not – like most deaf children to deaf parents – exposed to fluent ASL early on by deaf relatives form earlier generations. In addition, none of Simon’s deaf or hard-of-hearing classmates used ASL. The (hearing) teacher of the class communicated by using a manual code for English (which does not contain ASL morphology) simultaneously articulated with spoken English. The study examined Simon’s performance on a morphology task compared with the performance of his parents, as well as with eight children who had native signing parents. Due to the fact that Simon’s parents both learned ASL only as teenagers, their performance was characterized by many inconsistencies in particular in their use of ASL morphology. Nevertheless Simon’s production of ASL morphology showed to be (in most aspects) equal to that of children who were exposed to ASL from birth, and went beyond that of his parents. As pointed out by the authors, since nearly all children in normal linguistic environments are exposed to rich data for their primary language, 17 the question of whether children’s acquisition may proceed in uniform ways even in the absence of supportive linguistic input has been extremely difficult to study empirically. The constellation of Simon’s family is thus quite rare, but since Simon does not show signs of other perceptual, cognitive or social deprivation (like documented cases of feral children or children raised without human contact do), it constitutes an isolated case of extremely reduced language input.

Other studies of acquisition of language from reduced or degraded input (in otherwise socially well adjusted environments) have shown that profoundly

16 The reason for the parents learning ASL after the age of 15 was that their hearing parents first exposed them only to oral education at schools where no signing took place.
17 As discussed earlier, even input data under ordinary language acquisition circumstances has been characterized by proponents of the ‘Poverty of the stimulus’ argument as impoverished (mainly from the standpoint of the input presenting only a finite set of sentences of the language to be learned). But still, ordinary input data demonstrates a high degree of regularity as compared to cases (like Simon’s) when the input does not provide even the usual degree of consistency or complexity (Singleton & Newport, 2004).
deaf children educated through the oral method only (which relies upon lip reading and prohibits use of sign language) often develop their own gestural communication systems, so called homesign systems, that feature some of the properties of early child language (Goldin-Meadow & Feldman, 1977; Goldin-Meadow, 1978; 1982; 2003; Goldin-Meadow & Mylander, 1984; 1990; 1998).\textsuperscript{18} Also, children exposed to a pidgin\textsuperscript{19} as their native language create their own version (creole) of the language which shows greater complexity than the pidgin from which it was derived (Bickerton, 1981; Koopman & Lefebvre, 1981). On the other hand, since a similar type of structural expansion has also been observed in adult pidgin speakers (Mühlhäusler, 1979; 1980; 1986; Sankoff & Laberge, 1973; Sankoff, 1979), it has been argued to be a social phenomena related to the increasing demand for functional complexity that might not be unique to creolization in children. However, despite the ongoing debate on the source of the structural expansion, studies on homesign systems and creolization suggest that children may acquire a regular language system without input of such regularity (Singleton & Newport, 2004).

The acquisition of referential word function is closely related to conceptual behavior. Conceptual behavior is typically divided by psychologists into concept formation and concept utilization referring to the processes of learning a new concept and using an already known concept. Concept formation may – in view of the fact that some learning process often takes place before a concept exists – be thought of as a process more often engaging young or naïve organisms (Bourne, 1966). In contrast, utilizing concepts, perhaps more closely corresponding to adult behavior, often involves identifying or searching for one from a pool of familiar alternative concepts. However, since it is almost impossible to define when learning of a concept has been completed, and since completed learning does not have to be a prerequisite of using a concept, conceptual behavior in infants and adults is not clearly separable in conformity to the declared formation-utilization dichotomy. What does characterize onset of conceptual behavior in infants?

Since most adult concepts have verbal labels and some abstract concepts almost exclusively are used in verbal contexts, it is easy to forget that verbal labeling does not necessarily have to be a prerequisite of concept formation. The working hypothesis outlined in this thesis puts emphasis on the fact that even newborn infants’ weakly-established associations between different

\textsuperscript{18} Homesign systems demonstrate for example rather consistent gesture (word) order, use of embeddings (recursive structures), and gesture-internal morphology.

\textsuperscript{19} Pidgins are communication systems with reduced or simplified grammars that arise when speakers of many languages do not share a common language (Holm, 1988).
sensations may initially be called into use. In this view, precursors of conceptual learning in a broad sense may actually emerge prior to birth. For example, Varendi and colleagues (Varendi et al., 1996) suggest that infants are capable of prenatal olfactory learning. Varendi and colleagues demonstrated that human newborns are shortly after birth attracted to the odor of the maternal amniotic fluid reflecting the fetus exposure to that substance. In addition, it has been established that the fetus is able to perceive different tastes (e.g. a sweetener in the amniotic fluid) (De Snoo, 1937) and sights (a light blinking at mother’s abdomen) (Fulford et al., 2003; Johnson, 2002; Kiuchi et al., 2000; Eswaran et al., 2002). Similarly, the infant is here hypothesized also as capable of prenatal auditory learning. In fact, from about 150 days after gestation, fetal hearing is developed enough to show responses to outside noise (Lécanuet, 1998). From about 190 days growth the preborns’ cortex is assumed to be developed enough to be aware of and influenced by outside noise (Querleu et al., 1988). Thus prenatal auditory learning is here sketched as a detection process of contingencies between the alluded sensory representations available in the womb.

After birth, the infants’ preference for speech may be seen as consequence of a general recognition process. Therefore, rather than being a specific interest for speech as such, the same kind of recognition process applies to diverse types of acoustic stimuli to which the fetus might have been exposed (Lacerda et al., 2004). Furthermore, since the speech signal available to the fetus is dominated by low frequencies because of the surrounding tissue acting as a filter, it may be predicted that the newborn will focus on the same parts, mainly associated with prosodic and temporal aspects of the postnatal speech signal (Nazzi et al., 1998; Ramus et al., 2000). Also, the infant’s recognition of prosodic patterns of speech fits well with the type of $f_0$ modulation typically enhanced in infant-directed speech (IDS) (Kitamura & Burnham, 1998; Sundberg, 1998). In conclusion, instead of being an innate capacity, the newborns’ propensity towards speech and their remarkable learning power of language can also be explained as a consequence of prenatal learning. In addition, adults addressing infants tend to repeat words and phrases as well as intonation patterns (Papoušek & Hwang, 1991; Papoušek & Papoušek, 1989; Fernald et al., 1989). IDS may therefore additionally contribute to a supportive learning environment making specific motivation and explicit teaching superfluous. However, prenatal auditory learning influencing infants’ postnatal preferences for speech sounds in general and IDS in particular, as well as infants’ initial fuzzy (or incorrect) associations between sensory dimensions being guided by parental input might not be the whole story. Another reason for children’s conceptual behavior to have a lot in common with each other is that information processing is shaped by universal constraints on, for example, perception,
memory, and articulatory planning (to be discussed in Chapter 3 Distributional learning and in Chapter 4 Universal constraints on speech perception).

2.4 Perception of speech vs. non-speech

This section describes perception of speech vs. non-speech along two closely related subtopics. The first topic discusses whether language is processed by a domain-specific system (as proposed by Fodor’s modularity-concept and Spelke’s core knowledge-systems) or if processing of language is performed by a domain-general system. The latter view brings forward plasticity of the brain as support for the argument that the human mind might become modularized (as opposed to be prespecified in modules) as development proceeds. Further, a detailed explanation is presented of Bates’ functional infrastructure for language as a compelling model for describing general preconditions present at birth and how these skills might extend across the first three years of human life. The second subtopic in this section considers evidence from experimental studies which suggest that non-speech is not perceived in the same way as speech sounds are. The rest of the section will give an overview of research which in contrast has demonstrated existence of categorical perception for domains other than speech (e.g. perception of visual patterns). In particular, the experiment by Kluender and colleagues (1987) – in which the Japanese quail was shown to be able to categorize naturally produced speech stimuli – is explained in detail to illustrate that the ability to classify speech sounds might after all not be processed by a modular system of the human mind. The last paragraph of this section turns to the issue that it is the linguistic experience that affects our perception of speech sounds and that in some ways makes speech special as compared to non-speech.

Language is often thought of as so particular to its domain that it cannot be learned or processed by a general system. In favor of this position, Fodor’s (1983) modularity-conception suggests that the human mind is made up of genetically specified modules or input systems that are specialized for processing certain kind of input from the external environment (i.e. modules are domain-specific). Following this theme, Carey and Spelke (1996) have proposed that all our systematic knowledge is based on universal, preexisting representations of physics. Development of several complex cognitive skills, such as reading and calculation, are hypothesized to depend on core

---

20 The input systems encapsulate information which means only the module output (as opposed to internal processing) and information from lower-processing levels (but not from higher-processing levels such as top-down processes) are accessible as module input. Information is thus computed blindly by the input systems, and it is not until top-down hypothesis are derived that its general-purpose central processing starts.
knowledge-systems about agents, objects, space, number, motion, unity, persistence and identity.\textsuperscript{21} Each core knowledge-system represents only a subset of things that infants perceive (\textit{i.e.} core knowledge-systems are domain-specific) and functions to solve only a limited set of problems (\textit{i.e.} core knowledge-systems are task-specific) (Spelke, 2000).\textsuperscript{22} However, critics of Spelke’s theory have pointed out that it underestimates how much infants may learn through experience.

The strongest form of the constructivist position asserting that language must be learned and processed by a domain-general system was probably held by the Swiss developmental psychologist Piaget.\textsuperscript{23} As opposed to Spelke, Piaget argued that newborns are equipped with sensory capacities, but not with any real knowledge of, for example, object permanence. Approaching the modularity-generality debate slightly differently, Karmiloff-Smith (1996) points out that Fodor’s modularity-conception does not distinguish between prespecified modules and the process of modularization which, according to her, may occur “\textit{repeatedly as the product of development}” (p. 4). Thus, taking account of the plasticity of early brain development, Karmiloff-Smith puts forward that the mind may become modularized as development proceeds, and so argues for more epigenetic innately specified predispositions compared to Fodor: “\textit{Nature specifies initial biases or predispositions that channel attention to relevant environmental inputs, which in turn affect subsequent brain development}” (1996, p. 7).

The epigenetic view by Karmiloff-Smith closely resembles the interactionist view embraced by Bates. According to Bates, some functional infrastructures for language, starting with initial preconditions present at birth,\textsuperscript{24} are extended across the first three years of human life (Bates, 2004). Bates’ functional infrastructure of language is summarized in Figure 2.4.1. The first set of abilities (listed in the left-hand column) refers to the following properties:

\begin{itemize}
\item \textsuperscript{21} For example, newborns are assumed to have sense of object permanence (\textit{i.e.} apprehend that an object still exists even though it is no longer visible) (Kellman & Spelke, 1983), and to be able to orient in space by a natural sense of geometry-system that humans are endowed with (Landau et al., 1981; Landau et al., 1984).
\item \textsuperscript{22} Just like Fodor’s modules, the core knowledge-systems of Spelke are assumed to be encapsulated and to operate relatively independent from other cognitive systems. The flexibility of human cognition is achieved by combining representations from these isolated systems.
\item \textsuperscript{23} For overviews of Piaget’s theory see Piaget and Inhelder (1969), Belin (1989), Gruber and Voneche (1977), and Gold (1987).
\item \textsuperscript{24} According to Bates, these initial abilities (shown in the left-hand column in Figure 2.4.1) are general in two ways: they are not specific for language, or unique to humans. But nevertheless these skills are assumed to be exquisitely well-tuned in the human infant at birth or shortly thereafter.
\end{itemize}
(1) Object orientation refers to a fascination with small objects often manifested as a propensity to visually track, and to manipulate objects with hands and mouth.

(2) Social orientation refers to a strong tendency to orient towards social objects such as human faces and voices.

(3) Cross modal perception refers to a remarkable capacity to detect and contemplate spatial and temporal invariants in sound, vision, and tactile stimulation.

(4) Sensorimotor precision refers to a brain organized into sensory and motor maps working together with a surprising precision already at birth.

(5) Computational power refers to an impressive amount of computing power supporting rapid learning.

Figure 2.4.1: Functional infrastructure for language (adopted from Bates, 2004). The layout has been revised to include an arrow from imitation to secondary reasoning as explained by Bates.

The first set of initial skills is then, during the first 18 months of life, assumed to converge onto later developmental objectives (listed in the middle-column) as follows: (1) Orientation towards small objects and towards human beings is considered to lead to joint attention, which enables infants and their parents to focus and reflect on the same objects and events in the world. This ability is manifested through gaze following (by 3- to 9-months), pointing and production of deictic gestures (by 9- to 18-months) and positions the child in the right place for acquiring names for things. (2) Orientation towards human beings also intersects with cross modal perception permitting imitation which can take the form of reproduction of adult sounds, gestures and actions while attending to the shared objects. (3) Intersection of cross-modal perception and sensorimotor precision is considered to lead to sound-meaning mapping referring to infants’ ability to associate sounds with their meanings and (from about 1 year of age onwards) to pick up any signs, gestures, sounds or words that are reliably
associated with objects of interest. (4) The above skills together with the computational power of the brain appear to lead into rapid induction, *i.e.* “*a capacity for rapid statistical induction within and across domains*”\(^{25}\) (Bates, 2004, p. 251).

Finally, before the age of 36-months, the abilities in the middle column (Figure 2.4.1) are to be manifested in even more high-level cognitive behaviors (listed in the right-hand column) as follows: (1) The combination of joint reference and imitation are considered to lead to secondary reasoning, *i.e.* the ability to reason about the thought processes of other people also called theory of mind. (2) The same combination of joint reference and imitation appears as well to lead to observational learning, *i.e.* the ability to learn through silent observation of others. (3) The combination of sound-meaning mapping and rapid induction are sketched to lead into fast mapping, *i.e.* the ability to learn a new word in only one or two trials.

To summarize Bates’ model, a specific language learning-device does not exist. Instead, the entire human brain – quantitatively developing after birth leading to qualitative changes over time – is hypothesized to be a single general purpose learning-device (personal communication) permitting the acquisition of culture and technology as well as language.\(^{26}\)

Accumulated exposure to speech right from (and even prior to) birth, as well as use of speech for communication purposes make humans exceptionally experienced with the characteristics of speech as opposed to non-speech. However, disregarding for a moment that human speech perception is most likely strongly influenced by linguistic experience, non-speech sounds should be perceived in the same way as speech sounds for language to be processed by a domain-general system. Evidence for the opposite – modular speech perception – has been brought forward in classical experiments by Liberman and colleagues (Liberman et al., 1967). Specifically, Liberman’s *speech is special*-hypothesis holds that speech is perceived by a phonetic module which incorporates perception and production of speech. This biologically distinct module is assumed to be specialized to detect articulatory intentions of a human speaker (Liberman & Mattingly, 1985).

---

\(^{25}\) For examples for statistical learning mechanisms across domains see Chapter 3 Distributional learning.

\(^{26}\) Further, Bates maintains that all types of mental representations of new knowledge (compact, local, or broadly distributed) – irrespective of them being innate or learned – must be reflected somewhere in the brain in a form that is distinguishable from other aspects of knowledge. Therefore demonstrations that language is mediated by a particular part of the brain (localization) do not necessarily constitute evidence for innateness (Bates et al., 1998). For a theory of how nervous systems organize themselves, store information, and create new behavioral patterns see Edelman (1987).
Their speech is special-view originated from experiments on the existence of *categorical phoneme perception* (Liberman et al., 1967).

The notion of categorical phoneme perception is linked to the experimental demonstration of the fact that, if the second formant (F2) transition (carrying information about the place of production) of a synthetic speech signal is varied in acoustically equal steps through the range appropriate to produce /b/, /d/, and /g/, we do not hear gradual changes but jumps from one perceptual category to another. Thus, stimulus pairs differing by the same acoustic (F2) distance are perceived very differently depending on whether they represent distinct categories (one /b/ exemplar and one /d/ exemplar), or if they are variants of the same category. Mann and Liberman (1983) aimed at testing how transition cues are perceived in speech and non-speech. They removed the third (F3) transition cue from CV-stimuli varying along /da/-/ga/ continuum, and played the manipulated CV-stimuli and the excerpted F3 transition simultaneously one to each ear of the subject (*i.e.*, using dichotic listening-procedure). Listeners could correctly identify the syllables as belonging to /da/ or /ga/-categories, and like predicted by the notion of categorical phoneme perception, better discrimination of syllables at the category boundary was observed. In another variant of the experiment listeners identified the excerpted F3 transitions, sounding like chirping, as high or low. However, better discrimination of transitions at the (high-low) category boundary was *not* observed. Liberman and colleagues concluded that enhanced discrimination of speech sounds at category boundaries (categorical perception) *vs.* non-superior discrimination of chirping at category boundaries (non-categorical perception) indicated distinct mode of perception for speech and non-speech.

A similar mode of categorical perception (*i.e.* that one hears the phonemes but not the intraphonemic variations) has been found for /d/ and /t/ (Liberman et al., 1961b), /b/ and /p/ in intervocalic position (Liberman et al., 1961a), as well as for presence or absence of /p/ in /slit/ vs. /split/ (Bastian et al., 1959). On the contrary, Liberman and colleagues observed that if isolated (synthetic) vowels were varied in equal steps through a range to produce for example /i/ /ɪ/ and /ɛ/ these vowels were perceived to change along with the changes in the physical stimulus. The authors speculated that the steady-state vowels – as opposed to vowels in proper dynamic context – might not necessarily be perceived in the speech mode. Instead, Liberman and colleagues concluded based on evidence from Stevens (Stevens, 1966), that for vowels to be perceived categorically – and hence in the speech mode – the stimuli would have to be characterized by fast changes in formant position which is typical for vowels between consonants in rapid articulation.
However, these results have been challenged by experiments on the categorical perception phenomenon for visual patterns (Jusczyk et al., 1999), in prelinguistic infants (Eimas et al., 1978), chinchillas (Chinchilla laniger) (Kuhl & Miller, 1975; Kuhl & Miller, 1978), and quails (Coturnix coturnix) (Kluender et al., 1987). The existence of categorical perception for these indicate that (1) the categorization process is a component of several perceptual systems in humans, (2) knowledge of production constraints for phonetic segments does not have to be a prerequisite for categorical perception, (3) categorical perception might indeed be a component of the mammalian auditory system, and (4) that categorization of speech sounds, as opposed to requiring human-specific mechanisms, may work on a functional basis.

In greater detail, Kluender and colleagues (1987) demonstrated that Japanese quail categorized /b/, /d/, and /g/ followed by four different vowels in the training context and generalized what they had learned into test context. In the test context, the quails were capable of categorizing syllables in which the consonants /b/, /d/, and /g/ preceded eight novel vowels. Contrary to several previous experiments in which non-human animals have been tested on their discrimination or classification of synthetic speech sounds varying along certain dimension, the Japanese quail in the experiment by Kluender and colleagues were examined with respect to their ability to categorize naturally produced stimuli that varied on several acoustic dimensions. During training the three bird subjects received food reinforcement for pecking a lighted key during presentation of /d/ syllables, while they were required to refrain from pecking for ten seconds for presentation of /b/ or /g/ syllables to be terminated. During the test-phase 24 non-reinforced novel vowel trials were mixed with 96 non-novel vowel trials followed by the same reinforcement procedure as during training. In total eight test sessions were carried out two to three days apart. The results showed that all birds pecked significantly more often to novel syllables beginning with /d/ as compared with beginning with /b/ or /g/. The authors made a rigorous acoustic analysis (e.g. on voice onset time, second formant onset frequency, burst frequency and amplitude) of /d/ to answer for the birds’ correct categorization of new tokens. However, none of the analyzed properties could by itself support some invariant property of /d/. Therefore, the authors suggested that phonetic categorization in quails, just like visual

27 For a historical review of cross-language studies of speech perception see Strange (1995) and Eimas, Miller, and Jusczyk (1978).
28 The Japanese quail is a small (as compared with bobwhite quails) terrestrial bird which audiometric functions indicate close similarities with humans based on audiograms for frequencies below 5 kHz.
29 As pointed out by Kluender and colleagues, Hienz and colleagues (Hienz et al., 1981) have demonstrated that blackbirds and pigeons can make discrimination of steady-state vowels.
categorization in for example pigeons, might be based on combination of some available stimulus properties. In summary, research on auditory systems of non-human animals has confronted the theoretical claim by Repp and Liberman that phonetic categories and their “boundaries are determined by category prototypes that reflect typical productions of the relevant speech segments” (1987, p. 91). Such findings are at variance with the results of the Kuhl and Miller (1975; 1978) on chinchillas, and of Kluender and colleagues (1987) on quails which suggest that phonetic categories might be based on psychophysical discontinuities in the auditory system, as well as on purely functional grounds. These results thus question Liberman’s claim that speech sounds are perceived in terms of the underlying articulatory events, as well as positing uniquely human capacities in explaining speech perception.

Moreover, experiments designed to determine whether the McGurk effect holds for acoustically-specified non-speech events have found limited evidence for speech-specific processes. For example, when click-vowel syllables (that serve as consonants in some African languages) were presented to American listeners who perceive clicks as non-speech, a strong McGurk effect, comparable to those found for English syllables, was found (Brancazio et al., 2006). Among other studies suggesting that experience with a particular language affects perception of speech sounds, as well as that linguistic experience affects the perception of non-speech sounds, Johnson and Ralston (1990) report that perception of speech sounds is automatized, while the perception of less familiar sounds is not. Based on categorization and discrimination tasks for listeners reporting either speech or non-speech percepts, the authors conclude that “In terms of the "speech is special" debate, then, these results suggest that the perceptual processing of phonemes is distinct form the perceptual processing of other nonspeech events primarily by virtue of listeners' extensive experience in categorizing speech sounds” (Johnson & Ralston, 1990, p. 1).

2.5 Speech perception in typically vs. atypically developing infants

This section presents how word acquisition, according to the working hypothesis of this thesis, might be accomplished in blind and deaf infants. More explicitly, vision and hearing are regarded as typical precursors, but not necessarily prerequisites of conceptual development. In early communicative behavior, sighted or hearing parents are assumed to

30The McGurk effect, first described by McGurk and MacDonald (1976), is a perceptual phenomenon which illustrates the influence of synchronous visual articulatory information on perception of the acoustic signal. For example, a video of a person articulating /ga/ combined with a sound-recording of /ba/ is often perceived as /da/.

38
compensate for their infants’ impairment by using non-visual or non-auditory systems of communication to a greater extent. However, this section does not endeavor to give a detailed description of characteristics of developmental stages for blind and deaf children, but to provide a few examples of early compensatory communicative behavior observed in mother-infant dyads. Other substantive issues such as how experimental settings could be designed to allow for more appropriate evaluation of atypical language development are discussed next. Functional deficits that characterize language and cognitive syndromes in children with Down or Williams syndrome, children with diagnosis of specific language impairment (SLI)\(^3\), autism/autism spectrum disorder (ASD)\(^3\) are introduced. For example, deficits in social orientation, impaired categorization, impaired prototype formation and lack of central coherence are discussed.

The current working hypothesis pictures infants as capable of early word acquisition through exposure to multi-sensory information. In this view linguistic information is assumed to be derivable from implicit relationships between at least two types of sensory information available in the representation space of the infant. Characteristically, after some experience with the environment, typically developing infants establish sound-meaning associations based mainly on auditory and visual information. But how is word learning accomplished in children lacking one of these sensory inputs? Just like other living organisms which are exposed to a limited range of physical and biochemical properties available in their environment may become specialized within certain ecological niches\(^3\), the processing abilities and the type of sensory information available in the ecological settings of the infant are assumed to lead to exploration of mechanisms beneficial for language. Thus it is suggested that the typical language acquisition process – focusing on the acoustic signal as the most important factor – can be seen as a particular case of a general process capturing relations between different sensory dimensions (Lacerda & Sundberg, 2006).

---

\(^3\) For a review of discrepancy, inclusion, and exclusion criteria for language impairment, and specific language impairment, as well as for a discussion of the appropriateness of these diagnoses for research and clinical purposes, see Nettelbladt and Salameh (2007). Also, as pointed out by the authors, since language impairment is not a static condition, developmental changes should be taken into consideration when defining language impairments.

\(^3\) Autism is part of a larger diagnostic category, autism spectrum disorder (ASD). ASD include symptoms such as qualitatively reduced ability for social interaction and ability to communicate, and limited, repetitive and stereotype behavioral patterns, interests and activities (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, DSM-IV, by American Psychiatric Association).

\(^3\) Some bats exploring echolocation to survive in the ecological niche left open by their daylight competitors, and specialized bugs, such as the *Agonum quadripunctatum* or the *Denticollis borealis* (Nylin, 2006) that only emerge in the special environment created by the aftermath of forest fires are examples of living organisms specialized within certain ecological niches.
From this theoretical perspective what is then the developmental prediction for, for example, blind children? Or to put it slightly differently, if words are learned associatively, then blind children should have more initial difficulties to acquire words since they cannot identify objects that are not within reach (or follow the direction of what is being pointed at) while they hear words referring to the objects. Indeed, while taking into account the small size and heterogeneity of the blind and visually impaired population available for research (Ochaita, 1993; as cited in Puche-Navarro & Millán, 2007), initial referential development in blind children may be characterized as having some intricacies. For example, blind children (mainly under the age of 12-months) have initial difficulties in taking part in interpersonal communication and to share meanings (Price, 1991), as well as in assessing and getting the listener’s attention, and assessing and directing the listener’s attention to referents of their talk (Mulford, 1983). But nevertheless, by the age of three years, longitudinally studied blind children have been reported to perform within the typical range of sighted children on most language measures (Landau & Gleitman, 1985).

In the absence of the visual input – how is it possible for blind children to catch up the level of language competence of typically developing infants? A central issue here is that vision might be a typical precursor, but not necessarily a prerequisite of conceptual development. More specifically, as blind children may use tactile maps (i.e. mental maps containing tactile information) to acquire knowledge about spatial structure (Ungar, 2000), they are assumed to compensate for their impairment in auditory or verbal tasks by greater use of non-visual systems of communication (Burlingham, 1961; Mills, 1983; Andersen et al., 1984; Mulford, 1988). Also, infant-caregiver dyads with restricted access to visual communication have been shown to communicate in a refined way, using alternatives to visual communication (e.g. touch, vocalizations or facial orientation), already prior to the infants’ acquisition of language (Rattray & Zeedyk, 2005). For
example, in a longitudinal study on early communicative behavior of sighted mothers and their blind infants (Rowland, 1983), the mother’s vocalizations appeared likely to be followed by the infant’s smile, while the infant’s smile was most likely followed by the mother’s vocalization. In addition, mothers, sighted siblings and other caregivers certainly play an important role for the blind child as visual interpreters of the environment. And finally, difficulties experienced by blind children in the pre-verbal period, such as requesting displayed objects, are by some researchers believed to recede along with the development of verbal functions (Urwin, 1983). One way to interpret this is that in addition to conceptual development being integral to (or to precede) the emergence of words, some cognitive (conceptual) advances in blind children can, in line with the Sapir-Whorf hypothesis, be seen as a result of language development.

When infant and parent share the same modality for communication, language may be considered to develop through communicative interaction with a competent language user (Gallaway, 1998). However, already from the first weeks of life, as new born infants and their parents start to establish turn-taking, the communication process of deaf children is characterized by different strategies to those used by hearing infants. Contrary to hearing children who recognize familiar voices, the tone of voice used, and intonation cues signaling the end of turns (Spence & de Casper, 1987; de Casper & Fifer, 1980; de Casper & Spence, 1986; Nazzi et al., 2000), deaf children are more responsive to the parent’s touch and visual appearance.

In a longitudinal study (Urwin, 1978) sighted mothers were found in their communication directed to blind 15- to 22-month-old infants to prominently establish correspondences across different sensory systems by using their voices to match the babies’ facial expressions, and mark body movement. According to Urwin, the multiple accesses to and encoding of the environment by the mothers played a decisive role for these children to acquire language. Urwin suggested that studies on language acquisition in blind children should concentrate on relations between senses (as opposed to senses in isolation) and their potential substitutability.

The Sapir-Whorf hypothesis is a view about the relation between language and thought by Sapir (1884-1939) and Whorf (1897-1941). In its stronger version (linguistic determinism) the hypothesis states that language determines the way we think, while its weaker form (linguistic relativity) takes notice of that a particular concept in one language is referred to by a single word while another language requires several words to express the same thing. The fact that words for concepts in one language can be translated (and hence that the meaning of the concepts can be comprehended) into another language without words for the particular concept is often presented as the main argument against linguistic determinism. However, it is generally accepted that language may influence the way we perceive and remember objects and events: adults for example seem to recall things and make conceptual distinctions more easily if they correspond to words available in their language (Crystal, 1997).

Infant and parent share the same modality for communication for example when hearing children are exposed to spoken language by hearing parents, or when deaf children are exposed to sign language by deaf parents who sign.

I.e. the infant observes and responds to a communicative act initiated by the parent or the other way around. The roles of the participators of the dyad are then typically reversed.
The importance of the latter, and in particular the importance of eye-contact, becomes especially apparent in deaf parents’ efforts to facilitate the development of joint attention (*i.e.* the process whereby children and adults share a focus of interest about which communication takes place). For example, deaf parents frequently train their infants to visually track a manual point to a referent and then return eye gaze to the parent’s face, allowing communication about the referent (Kyle et al., 1989). Even when the child is not facing the parent, the parent tries to ensure that their signs are located within the child’s visual field. This is accomplished for example by adjusting the parent’s own body position relative to the child’s (*e.g.* moving in front of the child) or by signing on the child’s body (Herman, 2002). Successful joint attention is by some researchers subsequently assumed to lead to faster vocabulary acquisition both in hearing infants (Tomasello & Farrar, 1986; Akhtar et al., 1991), as well as in deaf infants (Harris et al., 1989). And finally, contrary to hearing infants who tend to increasingly start to vocalize in a speech-like way, the deaf infants – obviously due to lack of auditory feedback – follow this pattern to a very limited extent. Incidentally, some deaf children (often exposed to sign language from birth) producing manual babbling – *i.e.* sequences of hand gestures partially resembling meaningful, used signs (Petitto & Marentette, 1991; Petitto et al., 2001) – have been observed gradually to move towards usage of true signs found in the signing community.

The individual’s ability to represent the outside world is obviously mediated by the subject’s sensory system. Studies comparing performance of blind/visually impaired children to typical performance of sighted children often lead into considerations of poor or systematically backward functioning (Puche-Navarro & Millán, 2007), or to an interpretation of visual deterioration as a deficit (Pérez-Pereira & Conti-Ramsden, 1999). As a consequence, experimental settings to evaluate atypically developing children in comparison to typically developing peers should be designed carefully. For example, several studies of linguistic symptoms in adults with aphasia (Blackwell & Bates, 1995; Dick et al., 2001; Miyake et al., 1994) have shown that vulnerability observed in aphasic patients can in fact be induced in normal adults when testing under certain experimental conditions. Therefore studies in which typically developing children are asked to process words and sentences under stressed conditions can be useful to expose difficulties faced by impaired children (Bates, 2004). For example, deficits in morphology or syntax can be induced in normal children under

---

40 In addition, concurrent vocalizations (sometimes called vocal clashes) which have been frequently observed between hearing mothers and deaf babies (Gregory, 1984) indicate difficulties in establishing turn-taking as a basis for subsequent communicative interaction.
conditions, in which the visual or auditory stimuli are temporarily compressed, the signal-to-noise conditions are unfavorable, or the cognitive load is high, simulating the impaired children’s limited attention or working memory. The methods to study atypical performance are apparently important since they bear implications for the approach to diagnosis and intervention in communicative disorders.

Another reason to adopt the simulation of deficit-method for clinical populations such as late talkers, children with Down or Williams syndrome, children with diagnosis of SLI, or ASD is that the symptoms reflecting breakdowns in the normal system may increase our understanding of whether the causes of the impairment are located within or outside the language system (Bates, 2004). For example, research on children and adults with SLI, which has been proposed to be a (set of) genetically transmitted disorder(s), \(^{41}\) demonstrates correlations between deficits and a series of cognitive functions, such as aspects of attention, symbolic play, mental imagery, and detection of rapid sequences of sounds (Leonard, 1998). The fact that language-impaired individuals frequently demonstrate weaknesses in (or loss of) other high-level cognitive functions points in the direction of vastly distributed processing of language (Bates et al., 1998).

What kind of functional deficits characterize these language and cognitive syndromes? Several researchers have suggested that ASD may result from deficits in social orientation (Baron-Cohen et al., 1985; Happé & Frith, 1996), in impaired social motivation and/or joint attention \(^{42}\) (Bates, 2004), or in impaired theory of mind (i.e. the ability to represent and reason about mental states like feelings, beliefs, desires, emotions, and intentions, etc.) (Bloom, 2000). Likewise it has been speculated that other non-linguistic deficits, such as impaired cognitive capacity (computing power) might be held responsible for the language disorders found in children with Williams and Down Syndrome, and that impairment in some aspects of information processing (spectral, temporal, attention, and memory) might have serious consequences for language learning in children with SLI (Bates, 2004).

Based the observation that individuals with autism have difficulty integrating information and generalizing previously learned concepts into new situations (Klinger & Dawson, 2001) impaired categorization (i.e. the ability to form categories) has been suggested to be a potential functional deficit in children with autism. As maintained by Kuhl (1985) categorization of speech sounds

\(^{41}\) The genetic origin of SLI has so far not been proven, but SLI is likely to be seen in children whose parents or siblings have a history of language learning problems.

\(^{42}\) For explanation of these terms see The functional infrastructure for language-model by Bates in 2.4 Perception of speech vs. non-speech.
Chapter 2 Early word acquisition

presupposes an ability to discover differences between sound group A and sound group B, but also an ability to discover similarity among equivalent instances of sounds belonging to one of these categories (see also Kuhl, 1991; 2001). Therefore, to be able to categorize speech sounds (phonemes) infants need to ignore (allophonic) variation in the acoustic realization of sounds. This can be variation on the basis of the context, the syllable position, as well on differences in anatomy and speaking style of the speaker.\(^\text{43}\) Categorization is a cognitive process which not only allows individuals to organize speech sounds into groupings, but also to organize information in general (words, objects) into conceptual categories. Klinger and Dawson (2001) hypothesized that the difficulty to generalize information across settings in individuals with autism is linked to impaired prototype formation.\(^\text{44}\) In their study, children with autism, Down syndrome, and typical development were tested in two category learning tasks: one task that could be solved based on a set of rules (e.g. the presence of long feet for an imaginary animal category determined the category membership), and another task (prototype condition) that could not be solved based on a set of rules (the test stimuli were a mathematical average of the imaginary animals seen during familiarization). Their results showed that while all the groups could form a new category in the rule-based task, only the typically developing children were able to do it in the prototype task. The authors concluded that children with autism seem to be able to categorize information during concept learning but they might use a different, a rule-based strategy compared to typically developing children.\(^\text{45}\) In parallel to the impaired prototype formation hypothesis, Frith and Happé (Frith & Happé, 1994; Frith, 1989) have presented a lack of central coherence hypothesis. It refers to the inability to draw together diverse information to construct higher-level meaning in context in autism.\(^\text{46}\) However, research on conceptual impairment has brought forward a mixed picture (Molesworth et

\[^{43}\] Further, Kuhl (1985) concludes that the ability to recognize that sounds produced by different talkers are equivalent is fundamental to the infant’s own production of speech. If the infant cannot equate the sounds produced by other people with the ones they themselves produce, then imitation, as an important learning-to-speak component, would be missing.

\[^{44}\] A prototype is defined by Klinger and Dawson (2001) as a summary presentation or a holistic picture of a category as opposed to a list of rules for a category. Rosch (1973) first defined a prototype as the most central member or the first stimulus to be associated with a category. In addition, adults seem to weight prototypical information more heavily in categorization than recognition (Estes, 1986).

\[^{45}\] In addition, Klinger and Dawson (2001) speculated that since the type of rules (based on a single feature) used in their experiment were quite simple, it was possible that more complex rule-learning might be the missing component in the absence of prototype formation. Also, compared to the rule-based task, the prototype task might have been more difficult and required greater information processing demands.

\[^{46}\] To remember the essence of a story, while not being able to recall the authentic surface form is an example of the tendency to draw together diverse information to construct higher-level meaning in context in typical development (Frith & Happé, 1994).
al., 2005). Results from testing children with autism and Asperger syndrome on similar materials as Klinger and Dawson (2001) failed to support predictions of impaired prototype effect or impaired central coherence for both participant groups (Molesworth et al., 2005).

Prototype-based and exemplar-based descriptions are often seen as alternative accounts of high-level representations. While the controversy on whether conceptual knowledge in typically developing children is represented as an abstraction – prototype – of information (Posner & Keele, 1968; Rosch, 1978) or by individual category instances – exemplar-based (Medin & Schaffer, 1978; Nosofsky, 1988) falls outside the scope of this thesis, the assumptions of prototype vs. exemplar-based information processing should be considered in experimental design.

2.6 Speech perception in humans vs. non-human animals

In this section studies in which non-human primates are taught to communicate in human language are differentiated from studies that demonstrate common perceptual mechanisms in primates or distinguish human perception from perception in other animals. The latter are of interest here. The role of using stylized (as opposed to natural) speech materials in experiments with non-human primates is discussed then. Specifically regarding associative mechanisms in non-human primates, a study showing that rhesus monkeys may associate conspecific vocalizations and facial gestures, as well as a study showing that sea lions and simian primates are able to establish equivalence classes, are described. To put the suggestion that syntactic recursion distinguishes humans from non-human primates, some conflicting experimental evidence from cotton-top tamarins and starlings is presented. Finally, this section reviews some experiments reporting observations raising the question whether domestic dogs and chimpanzees may possess at least a rudimentary theory of mind.

A variety of communication (e.g. olfactory, auditory, tactile, visual, and vocal) among individuals is after all a common feature of several species that have social interaction at different levels. However human language stands out clearly among all the communication systems.

Several attempts have been made to teach non-human primates to communicate in a human language. To overcome anatomical differences between primate species – such as quantitative variations in the speech organs, the nervous control, and the brain organization – these attempts

---

47 The smaller mouth cavity in chimpanzees is less suited for production of vowel contrasts. The mouth cavity in humans is larger as a consequence of the gradual postnatal descent of the
have mainly been carried out with help of sign language (Gardner & Gardner, 1969; Terrace et al., 1979), keyboards of symbols (Savage-Rumbaugh et al., 1998) or plastic tokens standing for words (Premack & Premack, 1983). While most of the researchers agree that non-human primates are able to learn words, the extent of these animals’ ability to handle complex grammar is still at issue.

The aim of the current discussion is (1) to address the significance that common perceptual mechanisms in primates may have for human language and (2) to examine what characteristics distinguish human perception from perception in other animals. The first of these issues indicates that if a particular perceptual mechanism is found in humans and other primates then this shared mechanism might have been adapted for language processing in humans, demonstrating that the mechanism was initially not specific to language. Furthermore, non-human animal studies make it possible to separate effects of sensory representation from effects of learning. Thus non-human models may serve as ‘pure’ auditory systems, unadulterated by language experience (Kuhl & Miller, 1975; Kluender & Lotto, 1994). Studies of interest here investigate fundamental processes involved in hearing physiology, psychoacoustics, and perception in general, and association of auditory-visual information in particular. The second issue in

laryngeal skeleton relative to the hyoid and the descent of the hyoid relative to the mandible and cranial base (Crelin, 1973; Westhorpe, 1987; Lieberman et al., 2001).

The ability to maintain constant subglottal pressure necessary for production of continuous speech in chimpanzees is poorer. The control of the muscles between intercostals in humans is better as a consequence of the large size of the spinal cord (Jurmain et al., 1997). The control of these muscles typically allow for precise control of breathing during vocalization (Jones et al., 1992).

The activity level of the frontal lobe in chimpanzees is presumably lower. Corresponding activity in humans is assumed to be higher based on the disproportionately enlarged frontal lobe. The frontal lobe is regarded as the location of higher cognitive functions in humans (Deacon, 1997). Additionally, the control of fast or subtle movements of the tongue and lip muscles – especially needed for stop or fricative consonant production – in non-human primates is inferior. Corresponding control in humans is assumed to be better based on the enlarged proportion of the motor cortex involved in controlling these muscle movements.

It has been suggested that the upper limit of list of words reached by non-human primates corresponds to the level of about 2- to 2.5-year-old child (Hayes, 1953). Similarly, Savage-Rumbaugh and Fields (2000) have reported comprehension vocabularies of 500 words, and productive vocabularies of 150 words in bonobos.

While the three bonobos and one chimpanzee studied by Savage-Rumbaugh and colleagues (Savage-Rumbaugh et al., 1986; Savage-Rumbaugh et al., 1993) displayed sensitivity to the role of word order (i.e. their productions followed grammatical rules that reflect English word order), others have questioned these animals’ ability to understand how a change in word order may alter the meaning of a sentence. For example, as interpreted by Wynne (2004, p. 124), the chimpanzee called Kanzi (studied by Savage-Rumbaugh et al.) responded correctly only on twelve out of twenty-one pairs of reversible statements such as Put the ball on the hat, and Put the hat on the ball. And further, as criticized by Wynne (2004), Kanzi’s reactions to instructions were scored extremely generously (e.g. some commands had to be repeated several times with modifications, but Kanzi’s response was nevertheless scored as correct).
turn indicates that if humans alone are found to possess a certain perceptual or cognitive ability, then this skill must have evolved hand-in-hand with the human lineage and so might be crucial for human specialization in language. Consequently, cognitive skills in non-human primates (such as concept formation, abstraction, categorization, and learning) are of present interest.

One extreme position is to assume that there is no difference between the perceptual mechanisms of non-human primates and humans. For example, humans and chimpanzees are presumed to receive and process visual and auditory information in the same way in order to learn about statistical regularities in the world (Povinelli, 2000). In this vein, other non-human species, such as the Japanese quail (Kluender et al., 1987; Kluender & Lotto, 1994; Lotto et al., 1997) and the chinchilla (Kuhl & Miller, 1975; Kuhl & Miller, 1978) (as already discussed in 2.4 Perception of speech vs. non-speech) have proven to be effective models of human speech perception. Evidence from this type of non-human animal research has made it more problematic to maintain orthodox views about the existence of specialized speech perception mechanisms unique to humans, since non-humans demonstrate many of the speech perception phenomena that were initially described as uniquely human (Kluender et al., 1987; Liberman et al., 1967).

However, one thing that limits a number of these earlier studies in non-human animal speech perception is their reliance on highly-stylized speech tokens (for exceptions see e.g. Kluender et al., 1987). Typically experiments have employed synthetic phonemes, syllables or small clusters of syllables modeled after a single speaker. For this reason, our research group has recently started to use natural speech sounds with gerbils (Meriones unguiculatus) as a mammalian model of hearing and speech perception. Gerbils have excellent low-frequency hearing (Lay, 1972) which makes them a preferable animal for studies of audition within the human range of perception. Also gerbils are easily trained in classically-conditioned (Wetzel et al., 1998) and operantly-conditioned tasks (Sinnott et al., 1997). Their learning rates are much better than other species commonly used in speech

However, as pointed out by Bates and colleagues (1998), species-specificity does not necessarily constitute evidence for a specific mental organ for language. Instead, Bates and colleagues maintain that the human unique array of skills seem to be built on expansion of brain areas involved in higher cognitive functions, as well as in control of articulatory muscles.

For an exhaustive review of studies of cognitive skills in non-human primates see the encyclopedia by Tomasello and Call (1997).

As opposed to humans and monkeys who discriminate English vowels with about 100% accuracy, gerbils’ discrimination of very low-frequency distinctions in first formant (F1) frequency (e.g. /i/ vs. /ɪ/) seem to be inferior. However, since natural speech sounds have multiple distinguishing cues this should not be an issue.
Chapter 2 Early word acquisition

perception research (e.g. Japanese quail and chinchilla). In addition, the future intention of our research group is to shift focus from the phoneme level of speech perception to perception of full words in sentence context. For now, gerbils seem to be capable of discrimination (i.e. the ability to detect a difference between two sounds) of natural speech sounds (Holt, in preparation).

Investigating explicitly non-human primates’ ability to establish auditory-visual associations, Ghazanfar and Logothetis (2003) have shown that rhesus monkeys (*Macaca mulatta*) are able to recognize correspondence between their conspecific vocalizations and appropriate facial gestures. The coo-call of rhesus monkeys was typically associated with vocalizations characterized by small mouth opening and protruding lips, whereas their threat-call was mapped to vocalizations carried out with big mouth opening and no lip protrusion. Ghazanfar and Logothetis do not necessarily prove that primates are able to associate auditory-visual information as a basis of word-learning, but demonstrate that just like the perception of human speech, the perception of monkey calls (the auditory signal itself) may be enhanced by visual facial expressions. In an evolutionary perspective, bimodal perception in animals may be seen as a precursor of human’s ability to make multimodal associations necessary for speech perception (Ghazanfar & Logothetis, 2003). It was also demonstrated that rhesus monkeys are able to associate varying auditory signals with visual stimuli of varying shapes and colors (Gaffan & Harrison, 1991).55 The monkeys learned to associate each of six auditory stimuli (white noise, pure sine wave at 440Hz, a sequence of clicks, a sine wave with alternating frequency between 225-510Hz, synthesized word ‘bat’, and synthesized word ‘lick’) delivered from a loudspeaker, with one of six visual stimuli (constructed by superimposing one colored alphabetic shape upon a different alphabetic shape in a different color, e.g. a green Q on a large yellow X) displayed on a monitor screen. The visual stimuli on the screen were touched by the animal through the bars of the cage. The choice of the animal produced a repetition of the same sound, and (if the choice was correct) food. The results also showed that white noise was generally confused with other stimuli, while the words ‘bat’ and ‘lick’ were discriminated well from other stimuli, but confused with each other.

Associative learning mechanisms can account for novel and complex behaviors as shown by other experiments with animals as diverse as sea lions and simian primates (Schusterman et al., 2002). The animals in these

55 Gaffan and Harrison also investigated whether auditory-visual association may depend on a convergence of auditory and visual information in the prefrontal cortex of the animal. Rhesus monkeys’ ability to associate auditory and visual stimuli after ablation of some parts of the prefrontal cortex was severely impaired.
experiments have proved to be capable of perceiving dissimilar stimuli as if they constitute an equivalence class. More specifically, abstract problems were solved based on not only the animal’s capability to classify stimuli (e.g. signals or objects) along physical dimensions, but also according to their function. On the other hand, Dugdale and Lowe (2000) suggest that non-humans are not able to recognize stimulus equivalence. Their results showed that the chimpanzee subjects – who were trained to match stimulus A to B and then tested on their ability to match B to A – could not pass a test of symmetry of this kind. Many animals ranging from bees (Giurfa et al., 2001) to chimpanzees have, at least after intensive training under experimental conditions, demonstrated capability to understand sameness at some abstract level. In these match-to-sample experiments animals are taught to learn a rule such as ‘select the comparison stimulus that looks like the sample stimulus’. After learning this rule for one set of stimuli (e.g. color) the animals have been shown to be able to generalize it to another set of stimuli varying in another dimension (e.g. shape). Strictly speaking comparing human language communication behavior with the outcome of animal experiments has some unavoidable built-in caveats caused by the radically different nature of the experimental situations. Indeed, while human experiments inevitably call upon processing of ecologically relevant stimuli, non-human subjects are often forced to comply with artificial constraints induced by the experimenter.

Despite the fact that no other living species show human communication skills, interpretations which hold that non-human animals entirely lack cognitive abilities (i.e. Cartesian interpretations) are rare. Instead, the primate literature reveals that some foundations of the human conceptual system dealing with spatial, causal, and social reasoning are frequently present in other primates (Jackendoff & Pinker, 2005). Which important cognitive capacities might then separate humans from non-human primates? Hauser and colleagues (Hauser et al., 2002b) have initiated a current intense debate by suggesting that the capacity for syntactic recursion (as the only component of a uniquely human narrow faculty for language) differentiates human and animal communication. Consistent with this statement, cotton-top tamarins (Saguinus Oedipus) are shown to be able to spontaneously parse sequences (grammars) which can fully be specified by transition

---

56 Symbolicity is another frequently used term related to recognition of stimulus equivalence and generalization of sameness. For example, symbolicity has been used to refer to the potential inability of non-human primates to communicate about things that are not here and now or do not exist. According to this definition, human children are better at symbolicity since they learn new words for things and events they have never heard about or seen before simply by listening. Deacon (1997) maintains that the fact that children understand the relations between different symbols is the decisive factor that supports their understanding of the meanings.
probabilities between a finite number of states\(^{57}\) corresponding to words or calls (Hauser et al., 2002a), while monkeys tested with grammars at a higher level corresponding to human language\(^{58}\) failed to extract this type of hierarchical information (Fitch & Hauser, 2004). More specifically, in the experiment by Hauser and colleagues (Hauser et al., 2002a), tamarins were habituated to sequences of synthetic CV-syllables following the pattern AAB (e.g. wi wi di) or ABB (e.g. le we we). Subsequent habituation, the subjects were tested with two types of novel items, one of them following the pattern AAB and the other following the pattern ABB. The results showed that tamarins habituated to the AAB pattern were more likely to dishabituate to novel test items following the ABB pattern, while subjects habituated to the ABB pattern were more likely to dishabituate to novel AAB-pattern test items. Since experiments on human infants have demonstrated similar results (Saffran et al., 1996a), Hauser and colleagues (Hauser et al., 2002a) speculated that this mechanism shared by infants and tamarins might not have evolved specifically for the computational problems associated with language. Fitch and Hauser’s (2004) tamarins were instead habituated to CV-syllables (the A-syllables were spoken by a female and the B-syllables were spoken by a male, the female syllables are denoted in boldface) corresponding to either a FSG (AB)\(^n\) in which a random A-syllable was always followed by a single random B-syllable (and such pairs were repeated \(n\) times) (e.g. AB AB: no li ba pa), or corresponding to a PSG A\(^n\)B\(^n\) in which \(n\) sequential A-syllables were to be followed by precisely \(n\) B-syllables (e.g. AA BB: yo la pa do). The latter stimuli thus produced center-embedded constructions. Subsequently each monkey was tested with eight stimuli in random order. Four were novel stimuli consistent with the training grammar, and four were violations (but consistent with the other grammar). The results showed that tamarins mastered the simple alternating sequential pattern belonging to the FSG, but in contrast they failed to master the PSG in which components at one portion of a string were related to other components some distance away. However, these results have been challenged by Gentner and colleagues (Gentner et al., 2006) who showed that European starlings (Sturnus vulgaris) accurately recognized acoustic patterns defined by a recursive, self-embedding PSG. Species differences may thus be quantitative rather than qualitative distinctions (i.e. distinctions mapping onto formal grammars and automata theories) in cognitive mechanisms (Gentner et al., 2006). Further, as already pointed out by others (Pinker & Jackendoff, 2005), Gentner and colleagues speculated that the complexity and detail of human language compared to non-human

\(^{57}\) The finite state grammars (FSG) belong to the weakest class in the mathematical automata theory (the Chomsky hierarchy) of increasing generative power.

\(^{58}\) Sequences belonging to phrase structure grammars (PSG), also called context free grammars contain embeddings and long-distance dependencies.
communication systems might not be built on a single property such as capacity for syntactic recursion.

Learning a language might depend on the ability to extract and generalize regularities and hierarchical information, but if tamarins and songbirds like starlings have these abilities, why can’t they learn language? While these abilities may be necessary for language, most researchers hold that language learning in addition involves making assumptions about other peoples intentions and what they are talking about. Therefore, humans have been suggested to be unique in possessing theory of mind (i.e. the ability to represent and reason about the thought processes of others) (Povinelli & Vonk, 2004; Tomasello & Call, 1997; Povinelli & Eddy, 1996), although others suggest that chimpanzees have a rudimentary theory of mind (Hare et al., 2000; Hare et al., 2001; Premack & Woodruff, 1978). Povinelli and Vonk (2004) gave their chimpanzee subjects a choice to beg food from a person who could or could not see them. Their results showed that the chimpanzees begged for food from either person even though only the individual who could see the animal provided it with food. This indicated the chimpanzees’ inability to reason. On the contrary, Hare and Tomasello (1999) have showed that domestic dogs (Canis familiaris) seem to comprehend the implications of what they see others do: the dogs found hidden food as directed by the gaze of another dog or a person. Similarly, Hare and colleagues (2000; 2001) investigated theory of mind-potential in chimpanzees by placing two pieces of food between a high-ranking and a low-ranking chimpanzee. One piece was placed so that both could see it (and each could see that the other could see it), while the other piece of food was placed so that only the subordinate animal could see it (and the subordinate animal could see that the dominant animal could not see it). Their results showed that the subordinate chimpanzees preferred the piece of food that the dominant chimpanzee could not see, and therefore seemed to have at least a rudimentary theory of mind.

The relationship between human and non-human processing of speech sounds and representational ability is important for the theoretical outline of this thesis. Because infants can be assumed to be rather naive biological systems that eventually acquire their ambient language, they offer a unique opportunity to establish the extent to which their initial language development may be accounted for by general sensory representation mechanisms, also present in non-human species.
Part II Contributions of the present work
Chapter 3

Distributional learning

This chapter addresses distributional learning (or statistical learning of distributional information) within different domains in general, and within language learning in particular. Within language learning, distributional learning may be characterized as information derived from the relationships between linguistic units such as phonemes, morphemes or words. The role of integration of multiple probabilistic language cues (as opposed to distributional learning based on a single cue), and the relevance of distributional learning within natural language learning conditions (with reference to the structure of IDS) are discussed. Then, to demonstrate how processing of language input might be constrained by the working capacity (memory) of the infant, language learning in neural networks is exemplified. In addition to the background of earlier research on distributional learning of phonology, morphology, and grammatical regularities, a rationale for Study I and II is outlined expanding the scope of distributional learning to acquisition of words.

Previous work suggests that distributional learning plays an essential role in category formation across different areas of perception. For instance, there is evidence that distributional learning mechanisms operate on visual shapes in spatial (Fiser & Aslin, 2001; Younger, 1985) and temporal domains (Fiser & Aslin, 2002) as well as on tones (Saffran et al., 1999; Creel et al., 2004). In parallel to these studies, approaches to the study of statistical language learning have focused on the extent to which distributional learning mechanisms can capture different aspects of linguistic knowledge such as phonology, segmentation of the speech signal, morphology, and

59 For antecedents of statistical language learning see Charniak (1996).
60 Phonology refers to the phonological structure of the native language including the sounds that are legal in the language and in the way they are combined.
61 Segmentation of the sound patterns of language into words is complicated by the fact that in general in conversational speech there are no gaps or other obvious acoustic markers to signal the boundaries between words (Cole, 1980). In addition, it does not seem possible that segmentation is constrained by the lexicon (as suggested by the Cohort model by Marslen-
grammatical regularities (e.g. regarding word classes and elementary phrase structure). For example, experimental evidence suggests that infants between the age of 6- to 9-months are able to distinguish between prosodically similar, but phonologically different languages (e.g. English and Dutch) (Jusczyk et al., 1993). This effect disappears if phonological cues in the input language are eliminated by low-pass filtering the signal. Further, illustrating infants’ sensitivity to sequential structure of phonemes, Jusczyk and colleagues (Jusczyk et al., 1994) showed that 9-month-old infants listened longer to lists of words that contained frequent combinations of speech sounds. Other studies suggest that segmentation of the speech stream may be succeeded by predictability (or redundancy) as a distributional cue (Harris, 1955; Olivier, 1968; Wolff, 1977; Anderson, 1983). For example, both adults and 8-month-old infants are able to segment word-like units from a pseudo-language of concatenated nonsense words (Saffran et al., 1996a; Saffran et al., 1996b). This was accomplished after just 2 minutes of exposure to the continuous acoustic string. The cue that adults and infants appeared to be using was transitional probabilities between syllables. Syllables within words tend to have higher transitional probabilities (they are more likely to occur next to each another) than are syllables that span word boundaries. Saffran and colleagues speculated that listeners apparently use this tendency to segment continuous speech. 

However since statistical cues are probabilistic in their nature, it is relevant to ask to what extent adults and infants are in fact sensitive to the distributional properties of language. Although distributional statistics are useful at the syntactic level or for the purposes of speech segmentation, such distributional learning might not be the whole story. Indeed, since statistical cues are sometimes individually unreliable, integration of multiple probabilistic cues has been suggested to overcome the limitations of simple distributional learning (Christiansen, 2002). For example, if prosodic cues in

Wilson and Welsh, 1978) since the child may not know the words of the language until segmentation is successfully effected.

Morphology refers to the diverse ways (including prefixes, infixes, suffixes, etc. in regular and irregular forms) across languages to signal grammatical meanings such as past tense (Anderson 1992).

The infants’ (Jusczyk et al., 1994) sensitivity to sequential structure of phonemes was tested in a condition in which the subjects’ preference for frequent vs. infrequent combinations of speech sounds was measured.

In addition, Hauser and colleagues (Hauser et al., 2001) have demonstrated that cotton-top tamarins tested with compatible experiments, performed in very similar way in the segmentation task as human infants. Tamarins were first passively exposed to a stimulus set and then tested on isolated units. Their response was caught by whether they turned and looked toward a speaker hidden out of view. The methodology used did thus not require explicit training (or reward).
addition to phonological cues were present, both adults and infants were able to improve their performance still further (Saffran et al., 1996a; Saffran et al., 1996b).

Since distributional information is directly linked to the properties of the input, studies on heavily idealized materials (e.g. artificial data) may only be of limited use. Therefore, to explore learning mechanisms for detection and computation of distributional information under natural conditions, realistic input should be used. Artificial materials based on structures of IDS-corpora, such as within and across word-boundary transitional probabilities (Swingley, 2005) provide realistic input materials for children (Lacerda & Sundberg, 2006). Of course some simplifications (e.g. transcriptions at word or phonemic level) of corpora of natural languages are often necessary.

Distributional language learning is likely to be influenced by the fact that children are not able to attend to all statistical properties available in the input. In other words, there might a considerable difference between the existing information in the ambient language and the amount of information attended to, or received by the child. Elman (1993) suggests that limited processing abilities (corresponding to the limited memory span of the infant) in computational models of learning can describe the detection process of linguistic regularities in the input. To explore memory constraints involved in language users’ representation of grammatical embeddings at sentence level, a simple recurrent network was taught to process an artificial language. The task of the network was to predict the next word from the previous one in agreement with its categorical belongings and for example number agreement between subject noun and verb. In the first demonstration, when a full corpus of sentences of different complexities was used as training data, the network failed to master the task. In the second demonstration, when corpora of increasing complexity were used as training data, the network did well at the conclusion of training and generalised its performance to novel sentences. More specifically, the network was initially trained only with grammatically simple sentences, then with simple sentences and some complex sentences with more hierarchical relations, and finally with a corpus consisting of complex sentences only.

However, since training corpora of increasing complexity is not comparable to infants who hear exemplars of all aspects of adult language from the beginning, a third simulation was run. This time, the context units (i.e. the units that form the memory) of the network were initially set to random values after every two or three words, and then at increasingly long intervals. In other words, the temporal window within which the network could process information (corresponding to the working memory capacity of the infant) was first restricted to short sequences and later extended to longer
sequences. The learning results from this simulation showed to be as good as in the second demonstration.\textsuperscript{65} In summary, a system that starts with limited resources but develops more capacities over time seem to be at an advance in the sense that limited search space allows the network to solve a problem that can not be solved with adult like capacities.

The present dissertation studies I and II expand the scope of distributional language learning by addressing emergence of words. Infants are hypothesized to acquire words through sensitivity to statistical co-occurrences between auditory and visual information. To provide realistic input, the artificial speech materials used in these studies were based on the structure of Swedish IDS. The network simulations by Elman proposed that it is not the infant’s environment, but the infant that changes during the period (s)he is learning. In the current experiments with human infants, the subjects’ working capacity could obviously not be manipulated, but the materials used were adjusted in length and structure to compensate for the limited time for exposure available during experimental settings. The possible integration of multiple probabilistic linguistic cues in natural learning settings compared to the experiments is discussed.

3.1 Effects of linguistic variance, Design: Study I

In Study I infants ($N = 144$) in the age range of 3- to 20-months (mean age 12 months) were exposed to films displaying objects (puppets) while auditorily referring to the objects by target-words (names) embedded in phrases in an artificial language (Koponen, 1999). The aim of the study was to investigate to what extent infants manage to extract and associate novel (not previously heard) target-words to arbitrary objects. Two test conditions (low linguistic and high linguistic variance) in which the linguistic variance of the speech materials was systematically varied were created.

The artificial speech materials imitated the phonotactic, morphotactic, and syntactic structure of Swedish IDS, and were based on materials from a natural parent-infant setting. A twenty minutes long recording of a mother talking to her 3-month-old infant (Sundberg, 1998) was orthographically transcribed. Based on a description of phonotactics and morphotactics of these transcriptions, artificial words that correspond to different parts-of-speech (POS) were generated according to context-free word rules. Then, an analysis of performance during training showed that first, as the network was exposed to arbitrary input data, the internal representation space of the system was structured into distinct regions consisting of meaningful representations (e.g. for word categories, number, etc.). These representations then seemed to function as a notational vocabulary for learning the grammatical relations (understanding embeddings).
based on a description of syntax of the IDS materials, the words were embedded in declarative, imperative and interrogative phrases generated according to context-free phrase rules. A stop-lexicon based on a corpus of written and spoken Swedish was used to filter out semantically meaningful nonsense-generated words. To attain prosodic characteristics of IDS, the phrases were recorded while read aloud by a female speaker. Thus the properties of the artificial language (from now referred to as Svensiska) resembled the phonotactic, morphotactic, syntactic and prosodic structure of IDS while semantic cues to word-identity were eliminated. Compared to adult-directed speech, these materials were characterized by simplistic word initial clusters, few morphological derivations, and short (lexically) repetitive sentences.

The subjects \((N = 23)\) in the first test condition (high linguistic variance: Svensiska 1) were exposed to materials characterized by high linguistic variance: the target-word probability was low \((0.11\) out of total words), and the probability for a target-word to occur in the phrase-final position was low \((0.71)\). These materials strictly mimicked the structure of the natural IDS materials. Therefore, the materials contained context dependent references, for example pronouns – as opposed to target-nouns \(\text{per se}\) – were frequently used to refer to the objects. The target-words were presented in the initial, medial, or final phrase-position.

The subjects \((N = 118)\) in the second test condition (low linguistic variance: Svensiska 2 and 3) were exposed to materials characterized by low linguistic variance: the target-word probability was high \((0.17\) out of total words), and the probability for a target-word to occur in the phrase-final position was high \((1.0)\). Compared to the original IDS-materials of 20 minutes, the linguistic variance of the materials in this condition was reduced to meet the limited amount of time \((2\text{ min})\) for establishing sound-meaning associations during the experiment. Therefore, artificial target-words \(\text{per se}\) were used to refer to the objects. In each phrase, at minimum one repetition of the target-word was presented in the phrase-final position. In addition, some phrases contained supplementary repetitions of the target-word in the phrase initial or medial position. The materials in Svensiska 2 and 3 were identical, except that the presentation order of the puppets was revised.

The film materials (shown in Figure 3.1.1) consisted of an image (a sun) serving as an attention-catcher towards the screen and thereafter of four phases. Phase (1) baseline showed a split-screen of the two puppets (one to the left, the other to the right) during which infants’ spontaneous preference for the objects was measured. The audio was silent. Phase (2) presented one of the puppets slightly moving in the middle of the screen. The audio referred to the object by a target-word embedded in artificial declarative and
imperative phrases. Phase (3) presented the other puppet slightly moving in the middle of the screen. The audio referred to the object by another target-word in similar artificial phrases as during Phase 2. Phase (4) test-phase showed a similar split-screen of the objects as during the baseline. The audio referred to one of the puppets by a target-word embedded in artificial interrogative phrases. The materials were counterbalanced regarding left and right (during baseline and test-phase), and the presentation order of the puppets (Svensiska 2 vs. Svensiska 3).

Figure 3.1.1 The film materials consisted of an attention-catcher (a image of a sun) and the following phases: (1) a baseline showing a split screen of the two puppets (the audio was silent), (2) a presentation of one of the puppets acting in the middle of the screen (the audio referred to the object by a target-word embedded in artificial declarative and imperative phrases), (3) a presentation of the other puppet acting in the middle of the screen (the audio referred to the object by another target-word embedded in artificial phrases), and (4) a test-phase showing a split-screen of the two puppets (the audio referred to one of the puppets by a target-word embedded in artificial interrogative phrases).

The visual preference procedure, based on preferential listening procedure (Fernald, 1985), was used to measure infants’ looking behavior.

3.2 Results: Study I

The infants’ looking times were quantified as total looking time (s) at the target and non-target during baseline and target and non-target during test-phase (shown in the upper part of Figure 3.2.1) for the high variance-condition (Svensiska 1) and the low variance conditions (Svensiska 2 and 3). The net difference between looking time at the target and non-target during baseline (bias towards future target), and the net difference between looking time at the target and non-target during test-phase (post-exposure target dominance) are shown in the lower part of Figure 3.2.1.

In Svensiska 1, the non-target was shown to the left and the target was shown to the right during baseline, while the target was shown to the left and the non-target was shown to the right during test-phase. The bias towards future target (corresponding to the net difference between looking time at right_ab and left_ab in Figure 3.2.1) shows increased (above zero) looking time at the target during baseline, while the post-exposure target dominance (corresponding to the net difference between looking time at left_t and
right_t in Figure 3.2.1) shows decreased (below zero) looking time at the target during test-phase (shown in the lower part of Figure 3.2.1).

Figure 3.2.1 Mean looking times at target and non-target during baseline and test-phase in the high variance condition (Svensiska 1), and in the low variance conditions (Svensiska 2 and 3) are shown in the upper part of the figure. The materials were counterbalanced regarding left and right as follows: Svensiska 1 – target to right during baseline, and target to left during test-phase, Svensiska 2 – target to right during baseline, as well as during test, and Svensiska 3 – target to left during baseline, as well as during test-phase. The net difference between looking time at the target and non-target during baseline (bias towards future target) and the net difference between looking time at the target and non-target during test-phase (post-exposure target dominance) in Svensiska 1, 2, and 3 are shown in the lower part of the figure.

In Svensiska 2, the non-target was shown to the left and the target was shown to the right during baseline, as well as during test-phase. The bias towards future target (corresponding to the net difference between looking time at right_ab and left_ab in Figure 3.2.1) shows decreased (below zero) looking time at the target during baseline, while the post-exposure target dominance (corresponding to the net difference between looking time at right_t and left_t) shows increased (above zero) looking time at the target during test-phase.

Similarly, in Svensiska 3, the bias towards future target (corresponding to the net difference between looking time at left_ab and right_ab in Figure 3.2.1) shows increased (slightly above zero) looking time at the target during baseline, and the post-exposure target dominance (corresponding to the net
difference between looking time at left_t and right_t) shows increased (above zero) looking time at the target during test-phase.

The looking behavior of infants in the low variance condition (Svensiska 2 and 3) was – as indicated by the error bars – more stable compared to the looking behavior of infants in the high variance condition (Svensiska 1). Increased test-bias gain (the net difference between looking time at the target during test-phase and baseline) was taken as an indication of detected underlying sound-meaning association. A repeated measures ANOVA showed significant interaction between test-bias gain and test condition ($F(1,139)=9.734$, $p<0.002$) suggesting that the subjects’ looking behavior in the high variance condition (Svensiska 1) differed from the subjects’ looking behavior in the low variance conditions (Svensiska 2 and 3).

The looking time towards the puppets during exposure (Phase 2 and 3) was weakly correlated with the test-bias gain, indicating that longer exposure to the puppets was only slightly related to greater test-bias gain. The test-bias gain was not correlated with the subjects’ age or with the presentation order of the puppets (Svensiska 2 vs. Svensiska 3).

3.3 Evaluation: Study I

Studies investigating effects of serial position (primacy effect) present a mixed picture of children’s short-term memory. In some studies children seem to be better at encoding words at the beginning of a sentence (Newport et al., 1977), while other studies indicate more accurate encoding of words near the end of a sentence (Slobin, 1973). Similar biases has also been observed in adult subjects (Crowder, 1976). Therefore the materials in Study I were counterbalanced regarding the presentation order of the puppets. In addition, the materials were counterbalanced regarding left and right during baseline and test-phase (e.g. in Svensiska 1, the girl-puppet was shown to the right during baseline, while it was shown to the left during test-phase).

A prerequisite of performance in an auditory-visual task, like in Study I, is fully developed and contemporarily normally functioning hearing. Since the human auditory system as well as the neural structures to process speech are present and working adult-like already at birth (Kent 1997), the hearing abilities of the 3- to 20-month-old subjects in this study were assumed to be fully developed. Albeit not explicitly tested for the purposes of the current experiment, the actual hearing of the subjects was – based on parental observations and as revealed by Blicken orienterar efter ljud BOEL-distraction test, routinely used at Swedish health centers to screen infants 7- to 10-months – expected to be normally functioning (for a discussion on visual perception in infancy see 3.6 Evaluation Study II).
Analysis of the visual preference procedure materials (a close-up image video recording of the infant watching the film) was performed off-line frame-by-frame with a precision of 0.04 s (25 pictures per second). The experimenter labeled the subjects’ eye movements as directed towards left, right, front, or off-screen for the entire duration (3 min) of the film. The infant was seated on the parent’s lap so that the looking angle towards each of the objects was approximately ±15° from the perpendicular to the middle of the screen. Nowadays, frequently used infra-red eye-tracking systems based on corneal reflections, provide an automatic and more precise (with nominal precision of 0.02 s) method for tracking the infant’s gaze vector towards a test screen. Obviously such eye-tracking devices, compared to the visual preference procedure, allow for a much more flexible choice of areas of interest.

Despite of the relatively low accuracy for judgments of eye-movements, the visual preference procedure-materials can nevertheless be analyzed using different time-windows. In Study I, considering rather short duration of the baseline (30 s) and test-phase (30 s), total looking time at the target during baseline vs. total looking time at the target during test-phase were chosen as appropriate time periods for analyses. To capture the rather complex and rapidly varying looking behavior, the materials during test-phase could be analyzed using a short running time-window giving higher time resolution.

To determine existence vs. nonexistence of differences between the test conditions (high vs. low variance), a repeated measures analysis was used. An alternative to repeated measures is to use a multivariate analysis which statistically interprets the variation across levels of the independent variables as variance between subjects (Stevens, 1992) (Stevens, 1992). When making a decision about a null hypothesis (no difference between the means of the two groups on the dependent variable) four outcomes are possible: the researcher may correctly accept a true null hypothesis or reject a false null hypothesis, or the researcher may incorrectly reject a true null hypothesis (Type I error) or accept a false null hypothesis (Type II error) (Ventry & Schiavetti, 1986). Since the probability of making Type II error (falsely accepting a null hypothesis) increases with the multivariate approach, the current within-subject analysis was chosen.

Test directionality is another subject when evaluating the statistical analysis. Two-tailed (non-directional) tests are more strict or conservative than one-tailed (directional) tests meaning that a smaller difference may not be significant when analyzed with a two-tailed test, but may be significant as revealed by a one-tailed test. Thus, to the background of two-tailed tests often revealing in a sense more significant differences than those found with one-tailed tests, the hypothesized difference between the groups in this study
was in advance not suspected to be in one direction and therefore the more stringent two-tailed ANOVA was used.

3.4 The role of attribute types of objects, Design: Study II

In Study II infants \((N = 129)\) in the age range of 3- to 6-months were exposed to films displaying arbitrary geometrical objects while auditorily referring to them by artificial two-word phrases. The linguistic meaning of the artificial two-word phrases implicitly corresponded to the colors (red or yellow) and shapes (a cube or a ball) of the visual objects. In contrast to Study I, the aim of this study was to investigate infants’ ability to establish sound-meaning associations between properties of arbitrary objects and to them implicitly referring artificial words. Study II expanded the scope of emergence of sound-meaning associations by addressing attributes.

The subjects were divided in two age-groups. The infants in age-group I \((N = 53)\) were in the age range of 3- to 4.5-months (mean age 4 months), while the infants in age-group II \((N = 76)\) were in the age range of 4.5- to 6-months (mean age 5 months).

The film materials (shown in Figure 3.4.1) consisted of three phases. Phase (1) baseline showed a split-screen of the four geometric objects (upper-left, upper-right, lower-left and lower right) during which the infant’s spontaneous preference for the objects was measured. The audio played a lullaby to catch the infant’s attention towards the screen. Phase (2) presented four geometrical objects separately moving smoothly across the screen. The audio referred to the color and shape of each object by two-word artificial phrases. Each two-word phrase was repeated twice. Phase (3) test-phase showed an identical split-screen of the objects as during the baseline (Phase 1). The audio referred to one color and one shape of the objects by artificial target-words. Each target-word was repeated twice. Presentation of the objects (Phase 2) and test-phase (Phase 3) were repeated \(x 3\) times (from now referred to as exposure 1, 2, and 3, and test 1, 2, and 3). The colors presented during Phase 2 always appeared in novel shapes in Phase 3 (test-phase), and the shapes presented during Phase 2 always appeared in novel colors in Phase 3 (test-phase). For example, the color-yellow presented in the shape of cube and ball in Phase 2, appeared in the shape of cone in Phase 3, and the cube-shape presented in yellow and red in Phase 2, appeared in blue in Phase 3. The materials were counterbalanced regarding the presentation order of the objects throughout Phase 2, and regarding the choice of artificial words corresponding to the colors and shapes of the objects.
Figure 3.4.1 The film materials consisted of the following phases: (1) a baseline showing a split-screen of the four geometrical objects (the audio played a lullaby to catch the infant’s attention towards the screen), (2) a presentation of the four objects separately moving smoothly across the screen (the audio played two-word artificial phrases implicitly referring to the color and shape of each object), (3) a test-phase showing a split-screen of the four objects (the audio asked for one color and one shape of the objects by artificial target-words). Phase 2 (presentation) and Phrase 3 (test-phase) were repeated x 3 times. The colors presented during Phase 2 always appeared in novel shapes in Phase 3 (test-phase), and the shapes presented during Phase 2 always appeared in novel colors in Phase 3 (test-phase).

The (low intensity infra-red light) eye-tracking method was used to record infants’ eye movements.

3.5 Results: Study II

The infants’ looking times were quantified as normalized looking time at the target (red, cube, ball, yellow) during baseline (pre-exposure bias) and test-phase x 3 (post-exposure target dominance) for age-group I and II (shown in Figure 3, Study II, Part III Studies). For example, the value 0.4 on the y-axis indicates 40% of total looking time (sum of looking times at the four quadrants) during baseline (maximum 20 s), and the value 0.6 on the y-axis indicates 60% of total looking time (sum of looking time at the four quadrants) during one test-phase (maximum 2500 ms x 2).

Increased normalized looking time from baseline (pre-exposure bias) to test 1-3 (post-exposure target dominance) for target-shapes and target-colors was taken as an indication of detected underlying sound-meaning association. A repeated measures ANOVA (both age-groups) revealed significant deviations or tendencies (shown in grey in Table 3.5.1) from baseline to test 1-3 for cube and ball, from baseline to test 1 and 3 for red, and from baseline to test 2 for yellow, indicating that the subjects detected underlying sound-meaning associations reliably for shapes (cube and ball), but inconsistently for colors (red and yellow).

Pre-exposure bias to post-exposure target dominance contrast was correlated with the subjects’ age. A repeated measures ANOVA showed significant deviations or tendencies (shown in grey in Table 3.5.2) for age-group I (3- to 4.5-months): from baseline to test 1 and 3 for cube, and from baseline to test 2 for yellow, and for age-group II (4.5- to 6-months): from baseline to test 1-3 for cube, from baseline to test 1-2 for ball, and from baseline to test 1 and
3 for red, indicating that the older subjects detected underlying sound-meaning associations reliably for shapes (cube and ball), but inconsistently for colors (red and yellow), while the younger subjects demonstrate inconsistent capacity to associate recurrent visual shape/color dimensions with their implicit acoustic labels.

Table 3.5.1 Paired samples t-test (2-tailed, both age-groups). Significant deviations/tendencies from baseline (pre-exposure bias) to test 1, 2, and 3 (post-exposure target dominance) are shown in grey.

<table>
<thead>
<tr>
<th>Target, DoF</th>
<th>Baseline-test 1</th>
<th>Baseline-test 2</th>
<th>Baseline-test 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cube, (1,21)</td>
<td>F=21.660, p&lt;0.0005</td>
<td>F=12.875, p&lt;0.002</td>
<td>F=14.293, p&lt;0.001</td>
</tr>
<tr>
<td>Red, (1,18)</td>
<td>F=3.112, p&lt;0.095</td>
<td>F=0.949, p&lt;0.343</td>
<td>F=4.721, p&lt;0.043</td>
</tr>
<tr>
<td>Ball, (1,16)</td>
<td>F=6.457, p&lt;0.022</td>
<td>F=8.371, p&lt;0.011</td>
<td>F=3.861, p&lt;0.067</td>
</tr>
<tr>
<td>Yellow, (1,19)</td>
<td>F=1.121, p&lt;0.303</td>
<td>F=5.911, p&lt;0.025</td>
<td>F=2.906, p&lt;0.105</td>
</tr>
</tbody>
</table>

Table 3.5.2 Paired samples t-test (2-tailed) for age-group I (3- to 4.5-months) and II (4.5- to 6-months). Significant deviations/tendencies from baseline (pre-exposure bias) to test 1, 2, and 3 (post-exposure target dominance) are shown in grey.

<table>
<thead>
<tr>
<th>Target, DoF</th>
<th>Baseline-test 1</th>
<th>Baseline-test 2</th>
<th>Baseline-test 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cube (1,6)</td>
<td>F=15.075, p&lt;0.008</td>
<td>F=2.703, p&lt;0.151</td>
<td>F=4.898, p&lt;0.069</td>
</tr>
<tr>
<td>Red (1,7)</td>
<td>F=0.263, p&lt;0.624</td>
<td>F=0.508, p&lt;0.499</td>
<td>F=1.032, p&lt;0.344</td>
</tr>
<tr>
<td>Ball (1,7)</td>
<td>F=2.937, p&lt;0.130</td>
<td>F=1.143, p&lt;0.321</td>
<td>F=1.713, p&lt;0.232</td>
</tr>
<tr>
<td>Yellow (1,6)</td>
<td>F=0.860, p&lt;0.389</td>
<td>F=3.820, p&lt;0.098</td>
<td>F=1.104, p&lt;0.334</td>
</tr>
</tbody>
</table>

Total looking time towards the objects during exposure 1, 2, and 3 (Phase 2) (shown in Figure 3.5.1) was correlated with age-group (F(1,104)=2.849, p<0.094), indicating that while both age-groups’ looking times decreased towards the end of film, there was a tendency for the linear decline in looking time throughout exposure 1, 2, and 3 to differ in the two age-groups. This is indicated by the smaller error bars, and maintained higher looking time throughout the exposures 1-3 for age-group II.
Chapter 3 Distributional learning

Figure 3.5.1 Total looking times (sum of looking times at the presentation of four object in Phase 2) during exposure 1, 2, and 3 for age-group I and II.

Total looking time at target during test 1-3 was correlated with total looking time towards the objects during 3 x repetitions of exposure during Phase 2. A regression analysis showed significant correlations or tendencies (shown in grey in Table 3.5.3) for age-group I: between exposure and test 1-3 for cube, as well as between exposure and test 1-2 for red and ball, and for age-group II: between exposure and test 1-3 for cube, red, and ball, indicating that longer exposure to the materials may have led to longer looking times at target during test for both age-groups.

Table 3.5.3 Correlations between 3 x repetitions of exposure (Phase 2) and test 1, 2, and 3 (looking time at target) for age-group I and II (significant correlations/tendencies are shown in grey).

<table>
<thead>
<tr>
<th>Age-group I</th>
<th>Target</th>
<th>Exposure-test 1</th>
<th>Exposure-test 2</th>
<th>Exposure-test 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cube</td>
<td>r=0.37, p&lt;0.080</td>
<td>r=0.67, p&lt;0.001</td>
<td>r=0.54, p&lt;0.009</td>
<td></td>
</tr>
<tr>
<td>Red</td>
<td>r=0.41, p&lt;0.022</td>
<td>r=0.38, p&lt;0.059</td>
<td>r=0.06, p&lt;0.766</td>
<td></td>
</tr>
<tr>
<td>Ball</td>
<td>r=0.44, p&lt;0.022</td>
<td>r=0.42, p&lt;0.034</td>
<td>r=0.18, p&lt;0.418</td>
<td></td>
</tr>
<tr>
<td>Yellow</td>
<td>r=0.32, p&lt;0.166</td>
<td>r=0.09, p&lt;0.694</td>
<td>r=0.23, p&lt;0.348</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age-group II</th>
<th>Target</th>
<th>Exposure-test 1</th>
<th>Exposure-test 2</th>
<th>Exposure-test 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cube</td>
<td>r=0.57, p&lt;0.0005</td>
<td>r=0.36, p&lt;0.029</td>
<td>r=0.51, p&lt;0.002</td>
<td></td>
</tr>
<tr>
<td>Red</td>
<td>r=0.39, p&lt;0.038</td>
<td>r=0.39, p&lt;0.044</td>
<td>r=0.52, p&lt;0.006</td>
<td></td>
</tr>
<tr>
<td>Ball</td>
<td>r=0.31, p&lt;0.079</td>
<td>r=0.41, p&lt;0.018</td>
<td>r=0.52, p&lt;0.002</td>
<td></td>
</tr>
<tr>
<td>Yellow</td>
<td>r=0.35, p&lt;0.23</td>
<td>r=0.04, p&lt;0.819</td>
<td>r=0.14, p&lt;0.359</td>
<td></td>
</tr>
</tbody>
</table>
There were no significant interactions between the times spent looking at target and the choice of artificial words corresponding to the colors and shapes of the objects, or the object-position (left or right) on the split-screen. However, the subjects looked longer towards the upper quadrants red \((F(1,54)=37.564, p<0.0005)\), and cube \((F(1,63)=26.785, p<0.0005)\) than towards the lower quadrants during baseline.

### 3.6 Evaluation: Study II

The video materials in Study II were counterbalanced regarding the presentation order of the objects (during Phase 2), and the choice of artificial words corresponding to the colors and shapes of the objects. For example, the artificial words *nela* and *lame* were used to refer to either red or yellow, while the artificial words *dulle* and *bimma* were used to refer either to cube or ball.

While the target-words in Study I were embedded in artificial phrases based on the structure of IDS, the materials in Study II consisted of four novel artificial words. Therefore, the speech-materials in Study I resembled more the speech input faced by a young infant in natural settings. However, the task in Study II required generalization of the learned associations. In addition to recognition of target-objects indicating acquired context-bound knowledge in Study I, the infants in Study II were expected to generalize learned associations between target-words and colors/shapes into novel contexts (the colors presented in Phase 2 were displayed in novel shapes during test-phase, and the shapes presented in Phase 2 were displayed in novel colors during test-phase).

Study II investigated young (3- to 6-months) subjects’ ability to establish sound-meaning associations on shapes and colors of objects. Therefore infants’ visual development at this age is essential. While research has shown that infants’ ability to perceive size (Slater et al., 1990) and shape (Slater & Morison, 1985) is adult-equivalent already at birth, their visual ability is not considered to be comparable to an adult until the age of 3- to 4-months (Hainline, 1998). Infants’ visual acuity, binocular vision, and color vision are presumed to gradually develop during the first year of life. Thus, since little is known about the accurate starting point for infants’ color vision (as well as about the starting point for handling the concept of color as such), the infants in Study II were divided in two age-groups. The subjects’ actual visual ability was not examined, but before studying infants, the visual discriminability of the shapes and colors of the materials was tested and ascertained by a group of adults. A more rigorous design would include a similar visual-baseline test on infants.
In every experimental study, manipulation of the independent variable should show as a cause of a change in the dependent variable. One potential threat to internal validity is *time*, which may introduce differences (e.g., in terms of motivation, attention or interest) that take place within subjects in one age-group. To control for interference of such external factors, the looking behavior of the subjects (both in age-group I and II) was analyzed throughout 3 x repetitions of exposure and test. Presumably due to fatigue towards the end of the film, exposure x 3 towards the objects during Phase 2 was – both for age-group I and II – seldom correlated with total looking time at target during test 3.

When uttering an adjective-noun phrase such as *gul boll* (in English “yellow ball”), a native speaker of Swedish would typically place the sentence access either on *gul* when stressing the color or on *boll* when stressing the shape. In an attempt to reduce prosodic cues that could be provided by the naturally produced materials (read aloud by a female speaker), the component words of the two-word artificial phrases were recorded while they were read aloud in separate word lists (embedded in carrier phrases), and later concatenated to appear in a stress-neutral variant of each two-word phrase. An alternative way to prepare the materials would have been to let the sentence accent function as a cue to the target, in addition to already present phonotactic, morphotactic and syntactic cues. Using prosodic pointers as such, would probably lead to even stronger sound-meaning associations for attributes (Lacerda et al., 2005).

The infra-red eye-tracking method provides a recording of the infants’ eye movements with nominal precision of 0.02 s (50 pictures per second), and accuracy of 0.5°. The screen was split into four areas of interest for visual analyses corresponding to the four quadrants with origin in the center coordinates of the screen. The entire duration of the baseline (20 s), and a 2500 ms segment (starting 300 ms after the onset of a target-word) of the test-phase were chosen as appropriate time-windows for analyses (Fernald et al., 2001). An alternative way to analyze the looking behavior of infants would have been to measure their time point for shift in gaze toward target.

The choice of criteria for statistical analyses (two-tailed repeated measures ANOVA) in Study II were based on the same criteria as discussed in Study I.

### 3.7 Suggestions for further research based on Study I and II

The speech materials in Study I and II were words embedded in artificial phrases based on the phonotactics, morphotactics and syntax of Swedish IDS. As opposed to synthetic stimuli, the phrases were read aloud by a speaker to improve the naturalness of the materials. In Study I, the naturally produced speech materials also provided prosodic distributional information.
To investigate emergence of sound-meaning associations, the infants were tested with auditory materials in combination with visual materials. The visual materials were adapted for the special purposes of studying infants’ ability to associate target-nouns (names) to objects (Study I) and to associate target-adjectives to attributes (properties) of objects.

External experimental validity typically addresses issues such as to which extent, to which populations, circumstances, and measurement variables the found effect may be generalized. To test generality of distributional language learning, studies on infants’ ability to establish sound-meaning associations between (a) *verbs* and *actions*, (b) words belonging to *native vs. foreign* language and objects, (c) words and objects in combination with some *other* (*e.g.* olfactory, gustatory, or tactile) *sensory information*, (d) words and objects in an infant population with *atypical* development, (e) words and objects in combination with *parental* (or simulated) *guidance*, could be conducted. To provide information on the neurological base of language acquisition, behavioral studies could be complemented by brain-imaging studies. Also, in accordance with the aims of MILLE and CONTACT (for a description of the projects MILLE and CONTACT see Chapter 1 Aims and outline), similar experiments as performed with human infants are to be carried out with non-human animals, and to be simulated in mathematical and computational models. The advantage of experiments with non-humans and computer simulations is that they allow examination of questions that are impossible, unethical or extremely difficult to answer with human listeners.
Chapter 4

Universal constraints on speech perception

This chapter exemplifies certain universal constraints on speech production, followed by potential universal constraints on speech perception, such as consequences of exposure to the ambient language, predictive perception, and perceptual salience. To the background of earlier research on universal constraints on speech, a rational for Study III, investigating identification of words from partial phonetic information in young infants, and a rationale for Study IV, exploring effects of perceptual salience on emergence of auditory-visual associations, are outlined.

Transition from early vocalizations to phonologically organized speech was early on portrayed as a discontinuous process (Jakobson, 1968), or considered to be a maturationally driven route largely unaffected by the environment (Lenneberg, 1967). On the contrary, plenty of research after that (for an exhaustive review of phonological development in infancy see Vihman, 1996) has brought forward a picture of continuous vocal development, characterized by a gradual interaction between auditory and visual and motoric abilities. For example, visual contact (Mulford, 1988), and hearing yourself and others (Stoel-Gammon, 1988) are needed for vocal development to proceed typically and on time. Also, infants’ exploration of their phonetic work space appears not to be merely stochastic, but rather shaped by universal (non-linguistic) production constraints such as sensori-motor salience and energetics (Lindblom & Lacerda, 2006). Sensory-motor salience biases the selection of articulatory movements towards sensori-motorically salient ones – infants initially attend to more salient aspects and then gradually develop more detailed representations of the input. Energetics shapes articulatory actions towards energetically cheap movements – the infant’s jaw is constraint by general characteristics of resonators demonstrating that moving fast, as well as moving slowly is (measured as rate of energy use per time and distance) costly (Hoyot & Taylor, 1981). These production constraints are part of human prespecifications, but they are not specific to human language (Lindblom & Lacerda, 2006). Which universal constraints shape speech perception? In line with the universal

---

66 For a phylogenetically based theory of the origins and precursors of infant and adult speech production see (MacNeilage, in press).
speech production constraints, the present working hypothesis calls attention to perception constraints related to infant’s auditory capabilities, consequences of exposure to the ambient language, predictive perception and perceptual salience.

Since the infant’s initial auditory capacity is assumed to mediate similar sensory representations of the speech signals as the adult’s, potential differences in behavioral response patterns to speech stimuli may be attributed to higher-level factors rather than peripheral psychoacoustic constraints (Lacerda & Sundberg, 2006). One such a difference is brought forth by data on infant vowel production, adult-infant interaction and infant vowel perception (Eimas et al., 1971; Kuhl, 1985; Lacerda, 1992a; Lacerda, 1992b; Lacerda & Sundberg, 1996), which all converge towards enhanced vowel contrasts along the height dimension. Other experiments pointing at that language acquisition is affected by experience-based constraints further demonstrate how the ambient language may shift infants’ focus from general to more differentiated and language-bound discrimination ability (Polka & Werker, 1994; Polka & Bohn, 2003). To conclude, studies indicating that vowel height seem to play a dominant role in the organization of vowels, as well as studies on infants’ attunement from language-general to language-specific perceivers of speech, assume the infant’s general auditory process to gain linguistic content in the course of the language acquisition process.

The Hyper & Hypo (H&H) theory proposed by Lindblom (Lindblom, 1990) regards speech communication as an adaptive process constrained by signal-dependent and signal-independent properties of speech. The speaker makes hypothesis of the predictability of the message (based on the listener’s presumed familiarity with the topic discussed, the language and dialect spoken, etc.) and adapts the phonetic specifications in the speech signal (signal-dependent information) in accordance with this. The listener’s understanding of spoken language is accomplished by a parallel process of decoding the signal-dependent information and as guided by the listener’s linguistic and conceptual knowledge and by her/his expectations of the contents of the message (signal-independent information).

Based on that comprehension of spoken language requires rapid integration of signal-dependent information with knowledge of the world, continuous (i.e. on the basis of initial phonetic information from left to right) speech processing has been suggested to be central to this efficiency in adults (Cole & Jakimik, 1980). Also, Fernald, Swingley, and Pinto (2001) investigated 18

---

67 Skilled listeners are found to be able to process 10- to 15 phonemes per second in casual conversation (Cole & Jakimik, 1980).
and 21-month-old infants’ ability to recognize words without access to complete acoustic speech signal. The subjects’ task was to associate pictures with auditory-stimuli presented as whole words in intact form and as disrupted words (in which 300 ms of the word was heard). Their results revealed similar reaction times and accurate recognition of whole words, as well as disrupted words for both age-groups. Instead infants’ word processing ability was observed to be related to their lexical development – the children with greater productive vocabularies were more accurate in their recognition. Thus, in addition to adults, infants (towards the end of their second year) also seem to be able to rapidly integrate signal-dependent and signal-independent information.

Listening to speech is clearly not a matter of just listening. Under typical face to face speech communication situations the visual input provided by the speaker’s face conveys important information that is continuously added to the proper auditory information. Indeed the visual component is so strong that it clearly interferes with the auditory information whenever mismatches between the two sources of information occur (see the McGurk effect discussed in 2.4 Perception of speech vs. non-speech). A number of studies have addressed different aspects of this sort of auditory-visual speech perception in adult speakers (Fowler, 1996; Dekle et al., 1992; Green & Kuhl, 1991) as well as its potential ontogenetic origins (Kuhl & Meltzoff, 1982; 1984; 1996; Kuhl et al., 1991). For instance, Kuhl and Meltzoff (1982) showed that 18- to 20-week-old infants may pick up the correlation between acoustic and articulatory characteristics of speech sounds. In their experiment children were exposed to a split-screen displaying two faces articulating /a/ and /i/ respectively, and an audio signal consisting of one of these vowels. The infants looked significantly longer at the face matching the vowel sound thereby suggesting that auditory-visual integration in speech perception is present at early age. Also, Bahrick (2003) proposed that infants’ initial word acquisition might be supported by general synchronous multimodal information. In her study 5-month-old infants were tested on their ability to discriminate amodal phenomena such as rhythm or tempo. The subjects were tested in the three situations created by all the meaningful combinations of the video and audio tracks: (1) a regular presentation of the auditory-visual recording of a plastic hammer hitting a surface, (2) the video track without sound, and (3) the sound track without video. Changes in the rhythm were only detected by the infants who received the auditory-visual

68 The finding that 18 and 21 month-old infants’ word processing ability was related to their lexical development departs from the findings from an earlier study by Fernald, Pinto, Swingley, Wineberg and McRoberts (1998) indicating an age dependence on 15- to 24-month-old infants’ ability to identify partial words.
condition but it was observed that the older the children were, the less they needed to rely on the coherent auditory-visual information.

An alternative hypothesis for what guides individuals’ attention towards certain characteristics of auditory and visual information may be based on Stevens’ (Stevens, 1965) study on cross-modal matches between loudness and ten other perceptual continua, such as weight, warmth and light intensity. His results suggest that equal stimuli ratios produce equal sensory ratios across modalities, opening for a novel interpretation of Kuhl and Meltzoff’s (Kuhl & Meltzoff, 1982) findings. In terms of the present hypothesis, their findings might be explained by a cross-modality matching between the loudness of the speech stimuli (/a/ and /i/) and the visual prominence of the mouth openings corresponding to these sounds.

Although several studies (Newman, 2004; Fernald et al., 1998; Fernald et al., 2001) have shown that adults and children possess the ability to identify words from partial phonetic information, less is known about this ability in young infants. Therefore Study III investigates 12- to 16-month-old infants’ ability to identify spoken words on the basis of their initial sounds. In addition, the particular age-group was selected to bring light on the issue of suggested parallel development for prediction of other people’s goal-directed behaviour (Falck-Ytter et al., 2006) and other social competencies, such as communication (typically assumed to emerge around 8- to 12-months of age).

Study IV investigates infants’ ability to match auditory with visual information for speech and non-speech, as well as across speech and non-speech modalities. Study IV thus expanded the scope of detecting synchronous auditory-visual speech information by adding an investigation of the infants’ ability to perceive synchronous auditory-visual input for non-speech, and across modalities. The methodological improvement included using a high resolution eye-tracking system, which allowed the presentation of four images on a single screen during the test-phase instead of the two alternatives used by Kuhl and Meltzoff, thereby reducing the chance level of spontaneous looking at one of the images.
4.1 Prediction of phonetic information, Design: Study III

In Study III infants \((N = 15)\) in the age range of 12- to 16-months (mean age 13.7 months) were exposed to films displaying drawn images of familiar objects (a car, a watch, a teddy-bear, and a ball) while auditorily referring to them by stimuli presented as whole words in intact form and as disrupted (partly noise-replaced) spoken words. The aim of this study was to explore infants’ ability to predict phonetic information. Study III expanded the scope of predictive perception by addressing young infants’ ability to predict phonetic information.

The subjects were divided in two age-groups by the global median age. The infants in age-group I were in the age range of 12- to 13.5-months, while the infants in age-group II were in the age range of 13.5- to 16-months.

The target-nouns /klok:an/, (the watch), /bi:len/ (the car), /bol:en/ (the ball), and /nal:en/ (the teddy-bear) were selected according to frequency criteria based on the Swedish language assessment data base (Eriksson & Berglund, 1996) – a version of the MacArthur Communicative Development Inventories (CDI) based on parental reports. The target-nouns /klok:an/, (the watch) and /nal:en/ (the teddy-bear) were at all times presented as whole words in their intact forms. The target-nouns /bi:len/ (the car) and /bol:en/ (the ball) occurred solely in their manipulated versions prepared as follows: the first sound of /bi:len/ was extracted and concatenated with brown noise resulting in /b(i)/+NOISE (the /b/ carrying traces of /i/), and similarly the first sound of /bo:len/ was extracted and concatenated with brown noise resulting in /b(o)/+NOISE (the /b/ carrying traces of /o/). And finally, the same procedure was applied to the two first sounds of /bi:len/ and /bol:en/ – resulting in /bi/+NOISE and /bo/+NOISE.

The film materials (shown in Figure 4.1.1) consisted of three phases: Phase (1) baseline showed a split-screen of the four familiar objects (upper-left, upper-right, lower-left and lower right) during which the infant’s spontaneous preference for the objects was measured. The audio played a lullaby to catch the infant’s attention towards the screen. Phase (2) presented the first, second, third, and fourth familiar object. A recorded female voice named each object in carrier phrases, uttered in IDS style. The names/carrier phrases were repeated twice. Phase (3) test-phase showed an identical split-screen of the objects as during baseline (Phase 1). The audio asked for the watch (x 2 repetitions) and the teddy-bear (x 2 repetitions) by using intact words and for the car and the ball by using interrupted (in two versions) target-words.
The (Tobii 1750) eye-tracking system was used to record infants’ eye movements.

4.2 Results: Study III

The infants’ looking times (all subjects) were quantified as total looking time (ms) at each target and non-targets during baseline and test-phase (shown in Figure 2-5, Study III, Part III Studies). For example in Figure 2, the value 200 ms on the y-axis indicates looking time at the watch during baseline (maximum 20.000 ms), and the value 600 ms on the y-axis indicates looking time at the watch during the first repetition of the question asked (maximum 2000 ms).

Increased looking time from baseline to test 1 and 2 (the first and the second repetition of the questions on /klok:an/ and /nal:en/) for the intact target-words was taken as an indication of recognition of the familiar objects by their names, while increased looking time from baseline to test x 2 (the partially masked words /bi:len/ and /bol:en/ presented by their first sound only and by two first sound) was taken as an indication of infants ability to predict phonetic information. Paired samples (2-tailed, all subjects) test revealed significant departures/tendencies from baseline to test 1 for /klok:an/ (t(13)=2.482, p<0.027) and /nal:en/ (t(14)=2.057, p<0.059), for test 2 for /klok:an/ (t(12)=2.102, p<0.057), and for test 1 for /b(i)+NOISE/ (t(13)=1.814, p<0.093) (see Table 1, Study III, Part III Studies), indicating that intact target-nouns referring to objects were recognized reliably, while prediction of phonetic information (matching the phonetic information left in partially masked words onto their full lexical forms) was inconsistent.

Further analysis was based on the assumption that infants would spend 25% of their looking time on each of the objects during baseline. This assumption proved to be consistent with the actual distribution of the looking times.
during baseline, which showed a slight, non-significant, preference for the teddy-bear. The behaviour in the test-phase was then assessed in terms of departures from baseline and is expressed by the percentual increase (test gain percent) from expected looking time towards future target. Figure 4.2.1 shows results for each of the target-words grouped by age. Intact target-nouns /klok:an/ and /nal:en/ were recognized in a much more accurate way than the partially masked words bi:len/ and /bol:en/. It also shows that additional phonetic information (interrupted2) did not improve the infant’s ability to predict the target-word.

Figure 4.2.1 Relative gain (the net difference between looking time towards target during test-phase and expected looking time towards target during baseline) of looking towards target for intact vs. interrupted presentations of the target-word. Watch and Teddy-bear (shown in blue) were intact, while Car and Ball were partially masked and presented by their first sound (shown in green), or two first sounds (shown in yellow). The reference line indicates the relative gain assuming no looking preference for target in test-phase.

Age-group I (11.6- to 13.5-months) and age-group II (13.5- to 16.2-months) behaved in about the same way, although there was a slight advantage for the older infants in responding to intact words (see Figure 4.2.2).
4.3 Evaluation: Study III

To prepare the partially masked words, brown noise (containing energy at all the frequencies) was added to the first phoneme or the first two phonemes after excising the last phonemes of the signal. Using superimposed noise on the phonemes to be masked would on contrary require control of the class of phonemes to be masked and the energy levels of the masker at different frequencies for the masking effect to function effectively.

The first sound or the first two sounds of /bi:len/ and /bol:en/ were extracted and concatenated with noise resulting in /b(i)/+NOISE, /b(o)/+NOISE, /bi/+NOISE, and /bo/+NOISE. Just like the vowel of the syllables /bi/ and /bo/ was the cue to word identity, the initial /b/ (when only the first sound of the words was extracted) was carrying traces of the vowel and therefore functioned as a cue to /bi:len/ and /bol:en/. On the basis of literature on masking levels (Elliott, 1962), backward masking was assumed to be negligible.

The interrupted signals were not presented in a predictable or cognitively demanding manner. The infants’ spontaneous tendency to look at objects while hearing names referring to them was used.

The Fernald and colleagues’ (1998) study on 15-, 18-, and 24-month-old children’s ability for word processing revealed age dependence, while the study by Fernald and colleagues (2001) on 18- to 21-months old children indicated a lexical development dependence (the children with greater
productive vocabularies were more accurate in their recognition). Study III showed that subjects’ age was not clearly correlated with the Test-bias gain. Since the actual lexical development of the infants in Study III was not tested for, the hypothesis of lexical development dependence can not be addressed. However, the infants in both age-groups were younger than 16-months and therefore might yet not have entered the period known as the vocabulary spurt toward the end of the second year when the rate of learning new words typically begins to accelerate (Elman et al., 1996).

Just like in Study II, the screen in Study III was split into four areas of interest for visual analyses corresponding to the four quadrants with origin in the center coordinates of the screen. The entire duration of the baseline (20 s), and a 2000 ms segment (starting at the onset of a target-word \(X\) of each *Where is the X?* question) of the test-phase were chosen as appropriate time-windows for analyses (Fernald et al., 2001).

The choice of criteria for statistical analyses (two-tailed repeated measures ANOVA) in Study III were based on the same criteria as discussed earlier.

### 4.4 Detection of concurrence and synchrony in speech and non-speech, Design: Study IV

In Study IV infants \((N = 36)\) in the age range of 6- to 8-months (mean age 7 months) were exposed to films displaying (a) female actress articulating speech sounds (/ba/, /by/, /a/, /y/) while auditorily referring to *one* (/by/ or /a/) of them, (b) the actress clapping hands in different tempos (157%, 101%, 63% and 49% of the original recording tempo) while auditorily referring to *one* (101% of the original tempo) of the images by synchronized sound of hand-clapping, and (c) the actress articulating /by/ in different tempos (157%, 101%, 63% and 49% of the original recording tempo) while auditorily referring to *one* (101% of the original tempo) of the images by synchronized sound of hand-clapping. The aim of the study was to investigate infants’ ability to detect concurrent auditory and articularatory speech information, auditory non-speech and visual non-speech information synchronized in time, and auditory non-speech and articulatory information synchronized in time. Study IV expanded the scope of detecting concurrence between auditory and visual *speech* information by investigating infants’ ability to detect synchronous visual and auditory input for *non-speech* events, as well as *across* speech and non-speech *modalities*. Also, the spontaneous change level of looking at one of the images was reduced to 25% (as opposed to 50% change level) by presenting four (as opposed to two) alternatives on a single screen.
The film materials (shown in Figure 4.4.1) consisted of six phases: Phase (1) baseline showed a split-screen of four identical still images of the actress’s face (upper-left, upper-right, lower-left and lower right) during which the infant’s spontaneous preference for the quadrants was measured. The audio played a lullaby to catch the infants’ attention toward the screen. Phase (2) test-phase showed a split-screen of animated sequences of articulatory movements for /ba/, /by/, /a/, and /y/. The audio repeated /by/ (20 s), /a/ (20 s), /by/ (20 s) and /a/ (20 s) with rise-fall f0 contours. The target-image for /by/ and /a/ were placed in different quadrants for each of the four 20 s sequences. Phase (3) baseline showed a split-screen of animated sequences of the actress clapping hands in different tempos (upper-left, upper-right, lower-left and lower right) during which the infant’s spontaneous preference for the quadrants was measured. The audio played a lullaby. Phase (4) test-phase showed a split-screen of animated sequences of the actress clapping hands in different tempos (157%, 101%, 63%, and 49% of the original tempo). The audio played the sound of hand-clapping manipulated to 101% of the original tempo. Phase (5) baseline showed a split-screen of animated sequences of the actress articulating /by/ in different tempos (157%, 101%, 63%, and 49% of the original tempo) during which the infant’s spontaneous preference for the quadrants was measured. The played a lullaby. Phase (6) test-phase showed a split-screen of animated sequences of the actress articulating /by/ in different tempos (157%, 101%, 63%, and 49% of the original tempo). The audio played the sound of hand-clapping corresponding to the articulatory movements for /by/ in 101% of the original tempo.

Figure 4.4.1 The film materials consisted of the following phases: (1) a baseline showing a split-screen of four identical still images of the actress’s face (the audio played a lullaby to catch the infants’ attention towards the screen), (2) a test-phase showing a split-screen of four animated sequences of articulatory movements for /ba/, /by/, /a/, and /y/ (the audio repeated /by/ (20 s), /a/ (20 s), /by/ (20 s) and /a/ (20 s) while the placement of target was rotated for each sequence), (3) a baseline showing a split-screen of animated sequences of the actress clapping hands in different tempos (the audio played a lullaby), (4) a test-phase showing a split-screen
of animated sequences of the actress clapping hands in different tempos (the audio played the sound of hand-clapping corresponding to one of the sequences synchronized in time), (5) a baseline showing a split-screen of animated sequences of the actress articulating /by/ in different tempos (the audio played a lullaby), and (6) a test-phase showing a split-screen of animated sequences of the actress articulating /by/ in different tempos (the audio played the sound of hand-clapping corresponding to one of the sequences synchronized in time). The materials were counterbalanced regarding the presentation order of the speech and non-speech part of the films.

4.5 Results: Study IV

Analysis of the looking behaviour during baseline phases, and analysis (shown in Figure 6, Study IV, Part III Studies) based on percentage of the infants who, at a given time throughout each 20 s test-phase (with a running window of 200 ms), were looking towards each of the quadrants on the screen revealed a looking preference towards the upper quadrants, with a slight dominance for the upper right quadrant. To compensate for this bias, further results were sorted in terms targets being displayed, rather than the quadrants on which they appeared.

The infants’ looking times were quantified as the net difference (ms) between looking time at the target and non-targets during test-phase and looking time at the target and non-targets during baseline (shown in Figure 7-10, Study IV, Part III Studies). For example in Figure 7, 2400 on the y-axis indicates 2400 ms gain in looking time at /ba/ while listening to /by/.

Gain in looking time towards a target was taken as an indication of ability to detect (a) concurrent auditory and visual speech information, (b) synchronized sound of hand-clapping and hand-clapping movements, and (c) synchronized sound of hand-clapping and visual speech information.

A repeated measures ANOVA showed significant gain (a) for visual /ba/ while listening to /by/ (F(1,34)=5.243, p<0.028) and for visual /ba/ while listening to /a/ (F(1,34)=6.196, p<0.018) indicating that the infants looked at the visually most prominent stimuli, (b) for visual hand-clapping sequence in 101% of the original tempo while listening to the sound of hand-clapping in 101% of the original tempo (F(1,35)=3.987, p<0.054), indicating that the infants looked at the quadrant showing the clapping movements in synchrony with the sound, and (c) for visual /by/ in 157% of the original tempo while listening to the sound of hand-clapping in 101% of the original tempo (F(1,35)=9.758, p<0.004), indicating that the infants looked at the visually most prominent (fastest) stimuli.
Organized as a function of visual prominence, a linear trend in looking behavior towards /ba/, /by/, /a/, and /y/ while listening to /by/ (F(1,35)=7.235, p<0.011) and while listening to /a/ (F(1,35)=27.507, p<0.0005) in the speech condition was revealed. Also, when the looking behavior was organized as a function of the repetition tempo, a significant linear trend (F(1,35)=9.365, p<0.004) of increasing looking times for increasing frequency in the repetition of the articulatory movements in the cross-modal condition emerged.

When the sound of /by/ was matched to the visual image of /ba/, there was no significant interaction with the presentation order of the film materials, but the presentation order of the film materials was a significant between-subjects effect (F(1,34)=4.303, p<0.046), indicating that the performance of the group of infants who started the session seeing the speech stimuli was more stable than the performance of the group seeing first the non-speech sequences. No significant interaction or effect of the presentation order of the film materials was observed when the sound of /a/ was matched to the visual image of /ba/.

4.6 Evaluation: Study IV

The choice of the materials in Study IV was motivated as an extension of the original Kuhl and Meltzoff’s study. As a consequence the speech materials were not suitable for testing the cross-modality hypothesis inspired by Stevens’ Power law. To address this issue, new acoustic stimuli, involving direct loudness manipulations and visual stimuli of varying prominence are needed. In addition an objective measure of visual prominence, matching the loudness measures, will have to be used.

The video materials in Study IV were counterbalanced regarding the presentation order of the speech and non-speech sequences of the films. The placement of targets in the speech sequences (/by/ and /a/) was rotated for each 20 seconds sequence. In the non-speech sequences, looking bias towards the upper quadrants was compensated for by placing the targets (101% clapping tempo and 101% articulation rate for /by/) in the same quadrant during baseline and test-phase, and calculating gain in looking time by subtracting looking time at the future target during baseline from looking time at the target during test-phase.

The area of interest for visual analyses was defined within a radius of 100 mm from the centre of each of the quadrants. Since the sound of /by/ and /a/ were repeated throughout the test-phase, the entire duration of each test-phase (20 s) was chosen as appropriate time-window for analyses. The looking behavior in each test-phase was compared with looking behavior in
each baseline of the same length (20 s). Also, a running window of 200 ms was used to analyze infants looking towards each of the quadrants on the screen as a function of time during the entire duration of each test-phase.

The choice of criteria for statistical analyses (two-tailed repeated measures ANOVA) in Study IV were based on the same criteria as discussed earlier.

4.10 Suggestions for further research based on Study III and IV

In study III young infants’ ability to identify words from partial phonetic information was tested. In study IV infants ability to perceive synchronous visual and auditory input, both for speech and non-speech events was tested using a high resolution eye-tracking system allowing for testing with complex setting of visual stimuli.

To test generality of predictive perception, studies on (a) infants’ ability to predict phonetic information in interplay with some other (e.g. tactile) sensory information, (b) infants’ ability to predict non-speech information, (c) atypically developing infants’ ability to predict non-speech information, (d) non-human animals’ ability to predict others’ intention (like in the study by Falk-Ytter and colleagues with infants), and (e) non-human animals’ ability to predict phonetic information, could be conducted. In the same way, to test generality of the perceptual salience-phenomenon, other populations (e.g. atypically developing infants, or non-human animals) could be investigated.
Chapter 5

Discussion and conclusions

The present dissertation studies investigated the role of multisensory information as a precursor of sound-meaning associations in infancy. Specifically, the studies have addressed the questions of how linguistic variance (Study I), attribute type (Study II), and complete vs. incomplete phonetic information (Study III) affect infants’ ability to establish auditory-visual sound-meaning associations, as well as which characteristics of visual and auditory input affect infants’ ability to detect concurrence between synchronized speech and non-speech events (Study IV). Although the subjects of these studies were Swedish infants, the study results are not limited to perception in Swedish, but reveal potential universal aspects of speech perception. The present studies address distributional learning, predictive perceptual behavior, and perceptual salience as universal underlying mechanisms of speech perception.

It seems that the infants (3- to 20-months) in the low linguistic variance condition (in which the structure of natural IDS was compressed to compensate for the short experimental setting) in Study I could take advantage of grammar. They were capable of recognizing derivated, inflected and compounded stems of target-words which in turn were embedded in artificial declarative, interrogative and imperative phrases. Due to naturally produced phrases (that is read aloud by a female speaker), the inherent phonotactic, morphotactic, and syntactical regularities were supplemented with prosodic information (pauses, intonational patterns, etc.). Further, target-words were learned along circumstantial information provided by vision, presumably facilitated by frequently repeated artificial target-words presented in phrase-final position.

The results of Study II indicate that the infants (3- to 6-months) were able to learn covariance between target-shapes and artificial words. They also showed some evidence of being able to pick up the color references implicit in the materials. Longer exposure to the materials correlated with longer looking times at target for both age-groups. Further importantly, infants seemed to be able to generalize learned associations into new visual contexts.
The notion of distributional learning was expanded in Study I and II expanded from isolated speech sounds, or word extraction from connected speech to address emergence of sound-meaning associations. Sound-meaning associations that infants establish during natural settings are further linked to other consistent sensory (tactile, olfactory, etc.) information contributing to the contents of concepts, as suggested by the working hypothesis.

The infants (12- to 16-months) in Study III were able to recognize intact target-nouns, while prediction of partial phonetic information (i.e. matching the phonetic information left in partially masked words onto their full lexical forms) was inconsistent. The results did not support the idea that 12-month-old infants’ ability to predict other people’s action goals develops simultaneously with predictive phonetic perception. Since the subjects’ actual receptive and productive vocabulary size was not measured, recognition of incomplete acoustic signals can still be related to infants’ lexical development.

The infants’ (6- to 8-months) preferences in Study IV seemed to follow the salience of the visual articulatory displays (ba>by>a>y in the speech condition, and 157%>101%>63%>49% repetition frequency of /by/ articulations in the non-speech condition) on the quadrants. However they were able to pick up visual hand-clapping movements synchronized with the sound of hand-clapping. In this more complex experimental setting, the results did not support the notion that 6- to 8-month-old infants are able to pick up the correlation between acoustic and articulatory characteristics of speech sounds.

The investigation of predictive perception was in Study III expanded from studies on adults, school children, and older infants by addressing young infants. In natural settings the infants’ prediction skills of phonetic information are further supported by lexical development, in line with the working hypothesis. Study IV expanded the scope of detection of concurrence in speech by investigating infants’ ability to detect synchronous visual and auditory input for non-speech events, as well as across speech and non-speech modalities. The infants’ initially salience-based associative mechanisms which underlie the matching of auditory and visual information are presumably subsequently shaped by the infants’ growing experience of consistent events as they occur in natural settings.

Learning the meaning of words in spoken language typically requires association of information from the auditory channel with correlated information from the visual channel (or other sensory channels). Further, word acquisition was hypothesized to be part of a continuum ranging from fuzzy sensory associations in newborns, to simplistic sound-meaning
associations (6- to 12-months), and further to more abstract concepts (18- to 36-months). The relative success of the 3- to 20-month-old infants in Study I, and the 3- to 6-month-old infants in study II lends support to the view that early conceptual behavior is based on fuzzy associations between multi-sensory information. It also demonstrates that such initially fuzzy sensory associations may converge onto a rule-based capacity to associate sound patterns with visually consistent objects (as revealed by the correlation between post-exposure target dominance contrast and the subjects’ age in Study II).

After the relatively slow initial word-acquisition phase (until the infant's vocabulary contains about 25-50 words), the vocabulary spurt typically begins towards the end of the second year of life. However, based on an extensive parental-report investigation (The MacArthur Communicative Development Inventory) Bates and colleagues (Bates et al., 1994) observed that the age of entrance into the vocabulary spurt varied in between the end of the first year and the end of the third year of life. During the vocabulary spurt it was observed that up to approximately 100 words, nouns first increased in number, then between approximately 100 and 400 words, verbs increased mostly in number, and finally between approximately 400 and 700 words, function words increased mostly in number. Just like verbs, color-words are hypothesized to enter the child’s vocabulary after an increase of nouns in the lexicon. Therefore, to demonstrate a capacity to infer more abstract concepts such as colors (as compared to concepts like shapes which can be supported by tactile information), older infants need to be tested while the exposure to abstract concepts may have to be more extensive (e.g. several colors in several contexts) just like for extraction of the common denominator for verbs. However, the infants in Study II demonstrated at least a rudimentary ability to associate visual color dimensions with their implicit acoustic labels.

The fact that 6- to 8-month-old infants in Study IV failed to detect concurrent speech signal and articulatory information indicate that perception might be guided by a general match between salient auditory (e.g. high amplitude) and visual information (e.g. big mouth openings or lip movements) as suggested by Stevens power law. Also, the fact that infants failed to detect synchronized non-speech signal and articulatory information, but looked at the fastest articulation, instead indicates that infants match non-speech sounds to visually dominant cues (e.g. fast movements). And finally, the observation that infants were able to synchronize non-speech sounds to non-speech movements might depend on the fact that 6- to 8-month-old infants have sufficiently long experience with the ambient world to know that the sound of hand-clapping co-occurs with hand-clapping gestures. In terms of ETLA, these interpretations of the results give an ecologically relevant explanation for the infants’ behavior. Also, infants may
initially form non-relevant or ‘wrong’ associations, which will fade without reinforcement or interpretational value. Therefore the infants in Study IV might indeed at this age have focused on salient properties, rather than on stimulus properties that are ‘relevant’ according to an adult-definition. In natural settings, the infants’ ability to predict the environment will supposedly increase with the development of conceptual behavior of the infant.

The failure of the 12- to 16-month-old infants in Study III to predict phonetic information from disrupted words is an indication of the view that a substantial lexicon is needed to be able to process conversational speech continuously in an adult-like manner (Fernald et al., 2001). Based on the observation that the onset-age of vocabulary spurt varies between the end of the first and the end of the third year of life (Bates et al., 1994), it might be argued that the younger infants in Study III had presumably not reached the vocabulary spurt, while some of the oldest infants might have. Although the present infants’ vocabulary sizes were not measured, it can still be said that prediction of phonetic information is related to lexical development.

Postnatal speech development has, according to ETNA, precursors before birth, i.e. the fetus is capable of prenatal auditory learning. Although prenatal auditory learning was not tested in the present studies, it can help to explain some of the data, especially the successful establishment of sound-meaning connections in 3- to 20-month-old infants in the low linguistic variance condition (Study I) mimicking the lexically repetitive structure of IDS. This was presumably facilitated by prosodic cues and the phrase-final position of the target-words.

Research concepts such as the functional infrastructure for language by Bates (2004), the modularization view of Karmiloff- Smith (1996), the importance of starting small-hypothesis of Elman (1993), and the power law of Stevens (1965) are important to the hypotheses put forth here. As demonstrated by Elman, the infant’s limited focus of attention may effectively reduce the complexity of the ambient input during the early stages of language acquisition. Therefore, input to children does not have to be restricted – learning may occur as a result of limitations inherent the learning system.

Also, the infants’ inability to predict phonetic information, or that their perception is guided by salience instead of appropriateness, can be related to other possible interpretations of their immediate environment that are not necessarily in agreement with the adult’s perspective. These interpretations are comparable with the U-shaped learning exemplified by Plunkett and
Strömqvist (1992) (discussed in 2.2 Interdisciplinary perspective on language acquisition).

The present set of dissertation studies suggests that infants use associative mechanisms to acquire relevant linguistic information on the basis of multisensory regularities available in their immediate linguistic environment. Thus, the infant’s success in acquiring the relevant linguistic functions is in line with ETLA and may be more dependent on the structure of its linguistic environment than on the unfolding of a language acquisition program. Since the infant’s interaction with the ambient input must be taken seriously, much work on the role of multisensory information for language acquisition remains to be done. The intention of this chapter is to indicate directions in which further research might proceed.
References


Ref Type: Journal (Full)


Kyle, J. G., Woll, B., & Ackerman, J. (1989). *From gesture to sign and speech,* Final report to the Economic and Social Research Council (ESRC), Project No: C00232327 Bristol, UK: Centre for Deaf Studies, University of Bristol.


MacNeilage, P. The evolution of speech: A neodarwinian conception without linguistic nativism. (in press).


Sammanfattning på svenska


Nyckelord: språkinlärning; ordinlärning; audio-visuell; multisensorisk; förbindelser mellan ljud och betydelse; distributionell inlärning; perceptuell prominens; spädbarn
Part III Studies
Effects of linguistic variance on sound-meaning connections in early stages of language acquisition

Eeva Koponen, Lisa Gustafsson and Francisco Lacerda
Stockholm University, Sweden
E-mail: eevak@ling.su.se, lisag@ling.su.se, frasse@ling.su.se

ABSTRACT

To explore the processes underlying early sound-meaning connections an artificial language was created and used in a series of infant speech perception studies. The subjects were tested using the Visual Preference Procedure. The subjects’ responses were quantified in terms of looking times towards objects shown during the audio-visual exposure. Exposure to speech materials with large variance seemed to curtail the subjects’ ability to establish stable sound-meaning connections. However, reducing the linguistic variance led to successful sound-meaning connections. These results indicate that linguistic variance is one of the primary determinants of sound-meaning connections for 1-year old subjects. The paper will discuss how structural differences in natural language settings may account for the infant’s performance on word learning.

1. INTRODUCTION

Guttentag [1] suggests that the initial structuring of sensory input is an automatic process such that we simply connect what we hear with what we experience. More specifically, when coherent visual and auditory stimuli are stored they will eventually become associated [2]. Due to memory decay, less frequent sensory inputs are filtered out at the expense of more frequent ones [3], [4].

In the beginning of the ontogenetic development infants seem to have a limited working memory that presumably is able to process only simple linguistic structures [5]. But lack of variation in the linguistic input may cause the infant to make wrong generalizations. On the other hand, too much variation in the input may slow down the learning process until enough data are gathered to make correct generalizations [5]. With time, a more developed memory capacity will be able to handle larger variance in the input resulting in more economical processing. In evolutionary terms, emergence of grammatical structure can be seen as a result of processing an increasing amount of information. In this view syntax emerges under the pressure of expanding vocabulary [6].

Statistical learning mechanisms, or mechanisms which function like neural networks, have proved to be useful in segmentation tasks on synthetically produced syllables [7], [8]. In the present study we attempted to go beyond the acoustic signal per se to find out to what extent infants manage to extract words from continuous natural speech streams associating novel target-words with arbitrary objects. The hypothesis was that exposure to a nonsense language, characterized by the typical ‘repetitive’ structure of infant-directed speech (IDS: high target-word frequency rate, target-words in phrase final position), would be helpful in establishing correlations between the signal (speech sounds) and another co-occurring (visual) sensory input. Thus we assumed that statistical regularities conveyed by phonotactic and morphotactic constraints, part-of-speech (POS) and syntax may be helpful in making sound-meaning connections even for previously unheard, but well-formed utterances.

The reason for using an artificial language as a research tool in this study was thus to simulate a learning situation where the semantic content of words is arbitrary, and to control for subjects previous exposure. These artificial phrases, which were read aloud by a human speaker, were thus used to investigate infants’ sensitivity to linguistic stimuli.

Two test conditions were created: one control condition in which the structure of the artificial phrases was based on the structure of IDS, and one experimental condition in which the structure of the artificial phrases was manipulated to contain less linguistic variance as compared to the structure of IDS.

2. METHOD

The Visual Preference Procedure – a version of Preferential Listening Procedure [9] – was used. The audio-visual exposure consisted of 3 minutes long films. The objects (puppets) in the films were presented visually and auditorily with corresponding nonsense target-words embedded in nonsense phrases. A video camera recorded the infant watching the film. The infant’s looking behavior was later analyzed frame-by-frame.

2.1 SPEECH MATERIALS

The speech materials were based on a description of phonotactics, morphotactics, and syntax of Swedish IDS. A 20 minutes long recording of a mother interacting with her 3-month old infant [10] gave estimates of phoneme, morpheme and target-word frequencies, as well as typical proportion of declaratives, interrogatives and imperatives in Swedish IDS. The data collected was orthographically transcribed forming a mini corpus. As compared to Swedish adult directed speech, the word initial clusters in the mini corpus were rather simple, there were less morphological derivations and the phrases were short and characterized by lexical repetitions.
Context-free word rules were written to capture syllable structures of words corresponding to different POS in the infant-directed speech sample. Nonsense nouns (e.g. *bumann), pronouns (e.g. *bu), auxiliary verbs (e.g. *fur), verbs (e.g. *skrett), etc. were then randomly generated according to these rules. The nonsense words further corresponded to different structural constituents in declarative, interrogative and imperative phrases. Context-free phrase rules were written to capture these syntactical regularities. Finally nonsense phrases were randomly generated according to these rules (e.g. *Fur bu skrett bumann?, Swe. Har du sett nallen?, Eng. Have you seen the teddy bear?). To attain IDS typical modifications, such as frequent prosodic repetitions and expanded intonation contours, the phrases were read aloud by a human speaker.

2.2 DESIGN
Two conditions were created – one film with high linguistic variance (Svensiska 1) and another with low linguistic variance (Svensiska 2). To control for possible effects of the presentation order of visual materials another low variance film (Svensiska 3), consisting of exactly the same speech materials as in Svensiska 2, was created.

The high linguistic variance condition was characterized by many pronouns referring to objects, as well as target-words/pronouns occurring in other than phrase final positions. In contrast, the low linguistic variance condition was characterized by frequent target-word repetitions, the target-words always occurring in phrase final position. More specifically, in Svensiska 1 the probability for a word to be a target-word was low (0.11). About 70% of these target-words were in phrase final position and about 30% in other than phrase final position. In Svensiska 2 and 3 the probability for a word to be a target-word was high (0.17). These target-words were always in phrase final position. Consequently, the probability for a phrase to contain target-word in phrase final position was low (0.71) in Svensiska 1, and high (1.0) in Svensiska 2 and 3 (see Table 1).

<table>
<thead>
<tr>
<th></th>
<th>High variance condition</th>
<th>Low variance condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Film</td>
<td>Svensiska 1</td>
<td>Svensiska 2 &amp; Svensiska 3</td>
</tr>
<tr>
<td>Target-word probability (out of total words)</td>
<td>0.11</td>
<td>0.17</td>
</tr>
<tr>
<td>Probability for target-word to be in phrase final position</td>
<td>0.71</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 1: The probability for a word to be a target-word out of total words, and the probability for target-word to be in phrase final position in the high variance condition (Svensiska 1) are shown in the left column. The corresponding probabilities in the low variance condition (Svensiska 2 & 3) are shown in the right column.

In sum, Svensiska 1 strictly mimicked the structure of the natural IDS dyad. Instead of systematic presentations of target-words, the materials also contained pronouns and context dependent references, whereas the linguistic variance in Svensiska 2 (and 3) was reduced. Exposure to systematic target-words in phrase final position in Svensiska 2 (and 3) was assumed to facilitate the subjects’ establishing of sound-meaning connections during the experiment’s limited amount of time.

2.3 VISUAL MATERIALS
Each film (Svensiska 1, 2, & 3) consisted of four phases:

- Phase 1) showed a split-screen of two puppets: one to the left, the other to the right. Since the infants’ initial visual bias (spontaneous preference) towards future target was measured during this baseline phase, the audio of the phase was silent. Phase 1 lasted for 30 seconds.
- Phase 2) showed one of the puppets acting in the middle of the screen. The audio referred to the puppet with a target-word embedded in nonsense phrases. The infants’ looking time was taken as a measure of attention to this first puppet. This first exposure phase lasted for 60 seconds.
- Phase 3) showed the other puppet acting in the middle of the screen. The audio referred to the puppet this time with a new target-word embedded in similar nonsense phrases as in Phase 2. The infants’ total looking time was taken as a measure of attention to this second puppet. Also this second exposure phase lasted for 60 seconds.
- Phase 4) showed a split-screen of the two puppets again: one to the left, the other to the right. The audio referred to one of the puppets with a target-word embedded in nonsense questions. The infants’ total looking time at the target-object, as compared to bias towards future target in Phase 1, was taken as a measure of changed preference (Test-bias gain) indicating a sound-meaning connection. Phase 4 lasted for 30 seconds.

Svensiska 1, 2 and 3 differed regarding the presentation order (Phase 2 and 3) of the puppets.

2.4 SUBJECTS
The subjects were randomly selected from the National Swedish address register (SPAR) on the basis of age. A total of 144 subjects participated in the study (age range 102-604 days, mean age 360 days) (see Figure 1). Three subjects were excluded due to interrupted recordings (infant crying or not wanting to look at the film).

The remaining 141 subjects were randomly assigned to watch one, two or all three of the films: 23 to Svensiska 1 (age range 235-604 days, mean age 399 days), 118 to Svensiska 2 and 3 (age range 102-518 days, mean age 352 days). The parents participated voluntarily and were not paid for their participation.

2.5 PROCEDURE
A video camera recorded a close-up image of the infant
watching the film. The infant was seated on the parent’s lap. To reduce the risk of interacting with the infant, the parent listened to music through soundproof headphones during the whole procedure. Each infant’s recording was analyzed frame-by-frame (with precision 0.04 sec) as eye movements to left-, right-, front-, or off-screen.

Figure 1: The subjects’ age range varied in between 102-604 days (x-axis), mean age 360 days. A total of 144 subjects participated in the study (y-axis: subjects per age).

3. RESULTS

3.1 CHANGED PREFERENCE

The subjects in the low variance condition (Svensiska 2, and 3, N=118) showed increased mean looking preference at the correct object (longer Post-exposure target dominance than Bias towards future target), and the subjects in the high variance condition (Svensiska 1, N=23) showed decreased mean looking preference at the correct object (shorter Post-exposure target dominance than Bias towards future target). A repeated-measures ANOVA revealed significant interaction between the Test-bias gain (Post-exposure target dominance relative to Bias towards future target) and Condition (Svensiska 2, 3 vs. Svensiska 1) (F (1,139) = 9.734; p<0.002).

Figure 2: Mean looking times (Phase 1) at the future target, and mean looking times (Phase 4) at the target, in the low variance condition (Svensiska 2 & 3, N=118), and in the high variance condition (Svensiska 1, N=23).

3.2 EXPOSURE TIME

The results showed that an increased exposure (looking time) to the puppets (during Phase 2 and 3) was weakly correlated with the subjects’ changed preference: the longer the exposure to the puppets, the greater Test-bias gain (see the left dashed line in the middle of Figure 3). The Test-bias gain was calculated as the net difference between looking time at the target in Phase 4 (after presentation of the puppets) and initial bias (Bias towards future target) in Phase 1.

Figure 3: Exposure time to the puppets (during Phase 2 and 3) is shown on the X-axis (range 20-120 seconds), the age of the subjects is shown on the Z-axis (range 100-600 days), and the subjects’ changed preference (Test-bias gain) is shown on the Y-axis (range -20 to +20 seconds). Longer exposure time to the puppets was weakly correlated with greater Test-bias gain. The age of the subjects’ was not clearly correlated with the Test-bias gain. The subjects in Svensiska 1 are marked with triangles, and subjects in Svensiska 2 &3 are marked with circles.

3.3 AGE

The results showed that the subjects’ age was not clearly correlated with the Test-bias gain (see the right dashed line in the middle of Figure 3).

4. SUMMARY AND DISCUSSION

Recent work on infant speech perception has shown that concrete sounds are learned along circumstantial information about time and space [1]. Phonological, morphological, and syntactical constraints in the speech signal form statistical regularities [7], [8]. Establishment of audio-visual contingencies is, along with frequent visual input, benefited by these regularities [2]. Initially the limited working memory reduces the search space so that infants only need to entertain a small number of hypotheses about the world [11]. However, the dynamic nature of the learning process makes continuous modification of the infant’s hypothesis necessary as he/she is exposed to new information.

The aim of the present paper was to explore to what extent infants manage to extract and associate novel target-words to novel objects. Low target-word frequency rate, and varying target-word phrase position in the high variance
condition were strictly based on the structure of the natural IDS dyad that went on for 20 minutes. Thus the phrases in the high variance condition contained pronouns and context dependent references. The high target-word frequency, and constant target-word phrase final position in the low variance condition compressed the structure of the natural IDS dyad to better fit the experimental setting that went on for 3 minutes.

The results showed that the subjects in the low variance condition (Svensiska 2 & 3) had longer looking times at the target-object (Phase 4), as compared to Bias towards future target (Phase 1). This supports the notion that the nonsense phrases in the low variance condition could be associated with the objects just like semantically meaningful phrases do. Thus the infants’ establishment of sound-meaning connections was presumably supported by this low variance structure.

The shorter looking times at the target-object (Phase 4), as compared to Bias towards future target (Phase 1), in the high variance condition (Svensiska 1) may indicate the infants’ loss of interest as a consequence of too much variance during the exposure to the objects (Phase 2 and 3). As can be seen in Figure 2, the Bias towards future target in the high variance condition, probably due to more attractive puppets, was higher than the Bias towards future target in the low variance condition. This indicates that the infants in the high variance condition were indeed first interested in looking at the objects but their interest was apparently lost, perhaps because high variance exposure did not provide coherent enough information.

The subjects’ age range in this study was quite large (102-604 days). However age did not seem to be correlated with the Test-bias gain (the net difference between looking time at the target in Phase 4 and initial bias towards future target in Phase 1). When using the Test-bias gain as a measure of infants establishment of sound-meaning connections, one also need to take the exposure (looking time) to the objects (Phase 2, and 3) into consideration. The results showed that an increased exposure to the puppets was weakly correlated with the Test-bias gain.

It seems that the infants in this experiment could take advantage of grammar. They were capable of recognizing word stems embedded in derivated, inflected, and compounded nonsense nouns. The statistics inherent in the structure of phrases were, due to naturally produced nonsense phrases, supplemented by pauses and intonational patterns. In natural settings infants presumably benefit from syntactic and prosodic statistical information when learning different aspects of language. New experiments will be conducted in the future to test infants’ ability to establish sound-meaning connections between verbs and actions.

ACKNOWLEDGEMENTS

Swedish Research Council (Grant F0883/1999), Faculty of Humanities, and ERICSSON supported this research.

REFERENCES

Potential relevance of audio-visual integration in mammals for computational modeling

Eeva Klintfors and Francisco Lacerda

Department of Linguistics
Stockholm University, Stockholm, Sweden
eevak@ling.su.se

ABSTRACT
The purpose of this study was to examine typically developing infants’ integration of audio-visual sensory information as a fundamental process involved in early word learning. One hundred sixty pre-linguistic children were randomly assigned to watch one of four counterbalanced versions of audio-visual video sequences. The infants’ eye-movements were recorded and their looking behavior was analyzed throughout three repetitions of exposure-test-phases. The results indicate that the infants were able to learn covariance between shapes and colors of arbitrary geometrical objects and to them corresponding nonsense words. Implications of audio-visual integration in infants and in non-human animals for modeling within speech recognition systems, neural networks and robotics are discussed.

Index Terms: language acquisition, audio-visual, modeling

1. INTRODUCTION
This study is part of the “MILLE” project (Modeling Interactive Language Learning, supported by the Bank of Sweden Tercentenary Foundation), interdisciplinary research collaboration between three research groups – Department of Linguistics, Stockholm University (SU, Sweden), Department of Psychology, Carnegie Mellon University (CMU, USA) and Department of Speech, Music and Hearing, Royal Institute of Technology (KTH, Sweden). The general goals of the co-work are to investigate fundamental processes involved in language acquisition and to develop computational models to simulate these processes.

As a first step towards these goals, perceptual experiments with human infants were performed at the SU. These experiments were designed to investigate infants’ early word acquisition as an implicit learning process based on integration of multi-sensory (audio-visual) information. The current study specifically provides data on integration of arbitrary visual geometrical objects and “nonsense” words corresponding to attributes of these objects.

2. BACKGROUND

2.1 Research in mammals
As a theoretical starting point for the current study it is assumed that infants do not have a priori specified linguistic knowledge and that the acquisition of the ambient language is guided by general perception and memory processes. These general purpose mechanisms are assumed to lead to linguistic structure through learning of implicit regularities available in the ambient language. Thus, as opposed to the belief that initial innate guidance is a prerequisite for language acquisition [1], [2], the proposed Ecological Theory of Language Acquisition (ETLA) suggests that the early phases of the language acquisition processes are an emergent consequence of the interplay between the infant and its linguistic environment [3]. To be sure, observations based on implicit regularities may indeed lead to wrong assumptions and, just like other experience-based-learning, this type of trial and error use of words tends initially to create situated knowledge [4].

To be able to extract and organize implicit sensory information available from several modalities is an important ability for an organism’s success in its environment. Implicit learning processes are presumed to occur also in non-human mammals – a hypothesis being currently investigated by our collaborators at CMU. Explicitly concerning audio-visual integration in non-human animals – Ghazanfar & Logothetis [5] showed using preferential-looking technique that rhesus monkeys (Macaca mulatta) were able to recognize auditory-visual correspondence between their conspecific vocalizations (“coo” or “threat” calls) and appropriate facial gestures (small mouth opening/protruding lips or big mouth opening/no lip protrusion). Just like the perception of human speech, the perception of monkey calls may thereby be enhanced by a combination of auditory signals and visual facial expressions. Bimodal perception in animals was viewed by the authors as an evolutionary precursor of human’s ability to make multimodal associations necessary for speech perception. Such studies on non-human mammals are obviously important to examine questions that are impossible, unethical, or extremely difficult to answer with human listeners [6], like the pre and post operative comparisons of rhesus monkeys’ performance in auditory-visual (e.g. noise-shape pairs) association tasks supporting the notion that left prefrontal cortex plays a central role in integration of auditory and visual information [7].

2.2 Modeling within speech recognition systems, neural networks and robotics
After about one year’s exposure to their ambient language, children typically start to speak to interact with their environment. Despite of its complexity, children soon learn
the linguistic principles of their ambient language. However, this seemingly simple task is not easily transferred to formalized knowledge about language acquisition, nor is it easily integrated within speech recognition or operational models. Part of the problem that speech recognition systems battle with is presumably caused by the focus on the speech signal as the primary component of the speech communication process. Within natural speech communication speech is only one, albeit crucial, part of the process and the latest systems have started to make use of multimodal information to improve the systems’ communication efficiency. As an example, Salvi [8] analyzed the behavior of Incremental Model-Based Clustering on infant directed speech data in order to describe acquisition of phonetic classes by an infant. Salvi analyzed the effects of two factors within the model, namely the number of coefficients describing the signal and the frame length of the incremental clustering. His results showed that despite varying amount of clusters, the classifications obtained were essentially consistent. In addition to the model by Salvi the current co-work with the KTH further aims to develop a computational model able to handle information received from at least two different sensory dimensions.

We have recently reported [9] that two neural network models submitted to process encoded video materials that were earlier tested on a group of adult subjects in a simple inference task (similar to the task in the current study), performed well on both classification and generalization tasks. The task of the first type of architecture was simply to associate colors and shapes of the visual objects to the words corresponding to these two attributes, i.e. the model merely reproduced its input at the output level. The second architecture of the model attempted to simulate the adult subjects’ ability to apply their just learned concepts to new contexts. The performance of the model was tested with novel data not previously seen by the network (either new colors or new shapes). The performance of both network models was robust and mimicked the adults’ results well. Since the outcome of a neural network is dependent on the peculiarities of the input coding, its architecture and specific training constraints, the type of neural network models described here would presumably not mimic well enough the behavior of the infants in the current study. One reason for a vague match of behavior would probably be caused by the unlimited memory of the network which enables it to process data unrestricted as compared with infants’ restricted memory capacities. Also neural networks, often using sigmoid activation function for nodes, are non-linear regression models in which small differences in input value may cause large differences in neural computation behavior. Hence, a better way of modeling infant behavior would be to calculate effects of memory constraints when predicting audio-visual integration.

In addition to development of communicative ability leading to more sophisticated use of language, children quickly learn to interact with their environment through observation and imitation of manipulative gestures. To explore whether these motor abilities are developing independently, or if there are fundamentally similar mechanisms involved in development of perception and production of speech and manipulation of gestures, results from the current study and similar other experimental studies are tested within an embodied humanoid robot system. The hypothesis deals with a recent scientific finding pointing at the fact that action representations can be addressed not only during execution but also by the perception of action-related stimuli. Experiments with monkeys have shown that mirror neurons in premotor cortex (area F5) discharge both when a monkey makes a specific action and when it observes another individual who is making a similar action [10], [11] that is meaningful for the monkey. A mirror system is proposed also to exist in humans [12] and F5, the area for hand movements, is suggested to be the monkey homolog of Broca’s area, commonly thought of as an area for speech, in the human brain [13]. Both F5 and Broca’s area have the neural structures for controlling orolaryngeal, oro-facial, and brachio-manual movements [14]. Further studies in neuroscience suggest that there are parallels in mechanisms underlying actions such as to manipulate, tear, or put an object on a plate and mechanisms underlying speech actions recognized by their sound or mechanisms underlying speech systems [15].

For our research group at the SU the importance of modeling language acquisition lies unquestionably within the possibility of experimentally manipulating learning processes on the basis of experimental achievements and theoretical hypothesis formulated by us and other Neuroscience and Child-development partners.

3. METHOD

3.1 Subjects and procedure

The subjects were 160 Swedish infants randomly selected from the National Swedish address register (SPAR) on the basis of age and geographical criteria. Thirty-one infants were excluded from the study due to interrupted recordings (infant crying or technical problems). The remaining 129 infants were divided in two age groups: 26 boys and 27 girls in age-group I (age range 90-135 days, mean age 120 days) and 42 boys and 34 girls in age-group II (age range 136-180 days, mean age 156 days). The subjects were randomly assigned to watch one of four counterbalanced film sequences. The subjects and their parents were not paid to participate in the study. The infant was seated in a safe baby-chair or on the parent’s lap at approximately 60 cm from the screen. The parent listened to masking music through soundproof headphones during the whole session. The infants’ eye-movements were recorded with Tobii (1750, 17” TFT) Eye-tracking system using low-intensity infra-red light. The gaze estimation frequency was 50Hz, and accuracy 0.5 degrees. Software used for data storage was ClearView 2.2.0. The data was analyzed in Mathematica 5.0 and SPSS 14.0.

3.2 Materials

The films’ structure was: BASELINE (20 s), EXPOSURE1 (25 s), TEST1 (20 s), EXPOSURE2 (25 s), TEST2 (20 s), EXPOSURE3 (25 s), and TEST3 (20 s). The image used to measure infants’ pre-exposure bias in BASELINE is shown in Figure 1. During BASELINE the audio played a lullaby to catch the infant’s attention towards the screen.

The elements used in the EXPOSURE phases are shown in Figure 2. Each of the elements was shown in 6 s long film sequences. Each object moved smoothly across the screen while the audio played 2 × repetitions of two-word phrases, such as nela dulle (red cube), nela bimma (red ball), lame dulle (yellow cube), lame bimma (yellow ball), implicitly referring to the color and shape of the object. The two-word
phrases were read aloud by a female speaker of an artificial language in infant-directed speech style. The “nonsense” words were constructed according to the phonotactic and morphotactic rules of Swedish. However, the prosody of the phrases did not mimic Swedish prosody of two-word phrases – the words were instead pronounced as if they occurred in isolation, without sentence accent on either one of the words.

During TEST1, TEST2 and TEST3 the image shown in Figure 1 appeared again while questions such as *Vur bu skrett dulle?* (Have you seen the cube?) and *Vur bu skrett nela?* (Have you seen the red one?) were asked. The films were counterbalanced regarding choice of words corresponding to colors/shapes of the objects.

The results (Figure 3) showed increased looking times towards Red upper-left (UL), Cube upper-right (UR), Ball lower-left (LL) and Yellow lower-right (LR) from BASELINE to TEST1-3. The increments in looking time were larger in response to target shapes (Cube and Ball) as compared with target colors (Red and Yellow). This was in particular true for age-group II. Furthermore, repeated measures ANOVA on BASELINE to TEST1-3 contrasts revealed significant differences for Red, Cube and Ball.

The looking behavior of age-group II – as indicated by the error bars – was more stable than the looking behavior of age-group I. An analysis on BASELINE to TEST1-3 contrasts with age as a factor, revealed an overall age-group tendency for Red (*F* = 4.106, d.f. = 18, *P* < 0.058) indicating disparity in the looking behavior of infants in age-group I and II.

There were no significant interactions between the times spent looking at target and the choice of words corresponding to the colors and shapes of the objects or the object-position (left or right) on the split-screen. Infants did, however, look longer towards the upper quadrants (target Red *F* = 37.564, d.f. = 54, *P* < 0.0005 and target Cube *F* = 26.758, d.f. = 63, *P* < 0.0005) than towards the lower quadrants.

**4. RESULTS**

We predicted that if infants are capable of extracting the objects’ underlying properties, their looking times towards the relevant target color (Red or Yellow) and target shape (Cube or Ball) of an object will increase from BASELINE (Pre-exposure bias) to TEST1-3 (Post-exposure).

The looking behavior of age-group II – as indicated by the error bars – was more stable than the looking behavior of age-group I. An analysis on BASELINE to TEST1-3 contrasts with age as a factor, revealed an overall age-group tendency for Red (*F* = 4.106, d.f. = 18, *P* < 0.058) indicating disparity in the looking behavior of infants in age-group I and II.

**5. DISCUSSION**

Whereas early studies on language acquisition in infants were focused on production and perception of isolated speech sounds [16], recent experimental studies have addressed the structure of perceptual categories [17], and word extraction from connected speech [18], [19], [20], [21]. The current study further expands the scope of these investigations by addressing the emergence of referential function, integrating
recognition of patterns in auditory with the recognition of patterns in the visual input. In addition, our theoretical outline views early language acquisition as a consequence of general sensory and memory processes. Through these processes auditory representations of sound sequences are linked to co-occurring sensory stimuli and since spoken language is used to refer to objects and actions in the world, the implicit correlation between hearing words relating to objects and seeing (or feeling, smelling or otherwise perceiving) the referents can be expected to underline the acquisition of spoken language.

The materials in the current study were, as opposed to synthetic stimuli, naturally produced two-word phrases, whose linguistic meaning emerges from their situated implicit reference to the shapes and colors of the visual geometrical objects. The adjective-noun pattern in the two-word phrases followed the Swedish syntactical pattern. This situation, in which four new words were mapped onto a known reference to the shapes and colors of the visual geometrical phases required generalization of learned associations into visual integration during short exposure (about 1 minute) in young infants (< 6mos). Also, because success in the test phases required generalization of learned associations into new visual contexts, finding the correct visual target cannot be seen as simple “translation process”.

The age range of the subjects was 90-135 days (age-group I) and 136-180 days (age-group II). These age-groups were selected to investigate the extent of the age-related differences in the infants’ capacity to learn co-variance between the two different types of attribute targets (shapes and colors) and to them corresponding words. Indeed, despite of the fact that color words typically appear first after about ten months, these young infants showed some evidence of being able to pick up the color references implicit in the current experiment. Thus, although infants in this age range may not have the capacity to handle the concept of color as such, they at least demonstrate, as observed in other mammals, an underlying capacity to associate recurrent visual dimensions with their implicit acoustic labels.

6. ACKNOWLEDGEMENTS

Research supported by the Swedish Research Council (421-2001-4876), the Bank of Sweden Tercentenary Foundation (K2003-0867) and Birgit & Gad Rausings’s Foundation.

7. REFERENCES

Investigating the Emergence of Speech Communication – 
A Study on Infants’ Ability to Predict Phonetic Information

Klintfors Eeva (eevak@ling.su.se) 
Department of Linguistics, Universitetsvägen 10C, 
106 91 Stockholm, Sweden

Lacerda Francisco (frasse@ling.su.se) 
Department of Linguistics, Universitetsvägen 10C, 
106 91 Stockholm, Sweden

Marklund Ellen (ellen@ling.su.se) 
Department of Linguistics, Universitetsvägen 10C, 
106 91 Stockholm, Sweden

Abstract

The introduction of this paper provides an overview of infants’ prediction skills of action goals, as well as their ability to predict perceptual acoustic information. Prediction skills’ neurological correlates in general are discussed. A central hypothesis under investigation is that there are commonalities between the development of speech and manipulation. The current research is focused on the communication mode investigating infants’ ability to associate images of familiar objects with auditory-stimuli presented both as whole words in intact form and as disrupted (partly noise-replaced) spoken words. The looking behaviour of the infants’ was measured with the Tobii eye-tracking device. The results suggested that 11 to 16 month-old infants recognize the target object when the word referring to it was intact, i.e. when the name of the object was presented in its entirety. However, the infants did not seem to recognize the target object when the word referring to it was partially masked so that only its initial phonetic information was presented. These results indicate that young infants are sensitive to the phonetic information of the words and may need more extensive linguistic experience in order to derive full lexical forms from partially masked words. The paper concludes with suggestions for future demonstrations of infant anticipation of speech.

Introduction

Prediction of Other People’s Action Goals

Recent research indicates that one-year-old infants are able to predict other people’s action goals (Falck-Ytter et al., 2006). In these experiments, proactive goal-directed eye-movements of adults were compared with 6-month-old and 12-month-old infants’ gaze behaviour. The subjects were exposed to video presentations showing trials in which toys were moved by an actor’s hand into a bucket. The results showed that adults, as well as 12-month-old infants directed their gaze towards the bucket before the toy had reached it, while 6-month-old infants seemed to fail anticipating the goal of the action. The younger infants’ inability to predict the actor’s intention was suggested to indicate that predictive action perception in children develops simultaneously with other social competences such as imitation, “theory of mind” and communication. Such gestural and linguistic competencies are assumed to emerge typically by about 8-12 months of age. The authors also speculated on the possibility that the 6-month-old infants’ inability to predict the actor’s action goal could originate from their general inability to predict future events with their gaze. However this assumption was rejected as it conflicts with the results from earlier experiments demonstrating that 6-month-olds could indeed predict the reappearance of temporally occluded objects.

Falck-Ytter et al.’s (2006) study also suggested that successful goal prediction relies on observing the interaction between the hand of the agent and the object. Indeed, the results from another experimental condition where the objects moved along the same paths while the actor’s hand was not visible showed that predictive eye movements were not activated for any of the subject groups (adults, 12-month-old, or 6-month-old infants), suggesting that self-propelled objects are not perceived as performing goal-directed actions.

Neurological Coding System of Mirror-Neurons

The recent discovery of mirror-neurons presents a powerful neurological correlate of predictive perceptual behaviour. Activation of the mirror-neuron system (MNS) was first found in the ventral prefrontal cortex (area F5) of macaque monkeys (Rizzolatti & Arbib, 1998; Keysers et al., 2003) when a macaque observed another individual (human or monkey) performing an action that could be related to observer’s own repertoire of actions. An important aspect is that MNS was only activated when the goal of the action was clear to the animal. In fact, mimicking the gestures of goal-directed actions, without actually using the objects typically involved in the actions (like when pretending to peel an invisible banana) did not lead to activation of the MNS. The behavioural study by Falck-Ytter et al. (2006) is well in line with this assumption, indicating that the absence of an actor impairs the observer’s predictive perception of goal-oriented actions.
Furthermore, data from neuro-physiological and brain-imaging experiments indicates that MNS also exists in humans (Fadiga et al., 1995; Grafton et al., 1996) and it has been suggested that it may constitute a neurological base for coding empathy, social understanding and the ability for human communication (Rizzolatti & Craighero, 2004).

Mirror-Neuron System and Communication

Mirror neurons may account for coding of object-directed actions in which the gestural meaning is intrinsic to the gesture itself, but the question of whether the same system may be able to mediate abstract symbolic representations that are typically involved in human communication must also be addressed. In line with this, a relevant finding by Ferrari et al. (Ferrari et al., 2003) indicated that mirror-neurons of F5 may also code mouth-actions. Most of these “mouth mirror-neurons” were observed to become active both during the execution and observation of mouth-actions related to ingestive functions (e.g. grasping, sucking or breaking food) but some of the mirror neurons responded particularly to communicative mouth gestures (e.g. lip smacking). Therefore these findings extend the scope of MNS from hand-actions to mouth-actions suggesting that area F5, considered to be the homologue of human Broca’s area, is also involved in communicative behaviour.

From an evolutionary perspective, Ferrari et al. (Ferrari et al., 2003), suggested that understanding words related to mouth-actions may have evolved via activation of audio-visual mirror neurons. Mirror neurons initially responding to ingestive behaviour, may have led to a further development in which a set of F5 audio-visual mirror neurons eventually became responsive to the sound associated with the original actions, like hearing the sound of ripping a piece of paper without actually seeing the action (Kohler et al., 2002). Pursuing this evolutionary perspective Rizzolatti & Craighero (Rizzolatti & Craighero, 2004) speculated that the human individuals’ improved imitation capacities may have enabled the generation of onomatopoetic sounds without actually performing the action that originally generated the sounds. Hence, this capacity might have led to the acquisition of an auditory mirror system on the top of the original audio-visual one, progressively independent of it (see also (MacNeilage & Davis, 2000)).

Prediction of Phonetic Information

Earlier research (Samuel, 1996; Warren, 1970; Warren & Obusek, 1971; Warren & Warren, 1970) has shown that adults are able to interpret and reconstruct disrupted speech signals. In these studies adult listeners could identify both words disrupted by noise (Warren, 1970), as well as words disrupted by silence (Warren & Obusek, 1971). In the noise-disrupted case subjects reported that the disrupted words sounded intact. This phenomenon suggests that listeners perceive the word to continue behind the noise and “restore” missing phonemes – a phenomenon known as “phoneme restoration”. In the silence-disrupted case, subjects did not perceive the word as intact even though they were able to identify the word.

To examine school children’s perceptual ability in noisy environments, Newman (Newman, 2004) studied 5.5 year-old children’s ability to use partial phonetic information to identify familiar words. The results of this study showed that children’s perceptual ability, just like adults’, was better when speech signals were disrupted by noise as opposed to silence. However, compared to adults, children were overall more affected by signal disruptions, suggesting that young children are more dependent on the speech signal, particularly in noisy environments. These results are in agreement with Walley’s (Walley, 1988) who showed that children’s phoneme restoration demands more phonetic information than adults’ (i.e. children need to listen to longer portions of the disrupted word).

A study by Fernald, Swingley, & Pinto (Fernald et al., 2001) suggests that 18 and 21 months-old children are able to recognize words without access to complete acoustic speech signals. In this study the subjects associated pictures with auditory-stimuli presented both as whole words in intact form and as disrupted words in which only the initial 300 ms of the word was heard. Their results showed that children from both age groups could recognize whole words, as well as disrupted words. There were no differences in the two age-groups reaction times to these two types of stimuli. Instead their study indicated that children’s word processing ability was related to their lexical development – the children with greater productive vocabularies were more accurate in their recognition, a conclusion that departs somewhat from earlier findings by Fernald, Pinto, Swingley, Wineberg & McRoberts (Fernald et al., 1998) indicating an age dependence on the ability to identify partial words for infants in the age range 15 to 24 months.

The Nature of Infants’ Lexical Representations

There is an ongoing discussion on whether infants’ lexical representations are of a holistic or of a more specific nature. As an example, the fact that younger infants were not so accurate in identifying partial words was taken as an indication of lacking segmental structure in their lexical representations of words (Fernald et al., 1998). On the other hand, studies concerning language-specific tuning of vowels have shown that 11-12 months-old infants are sensitive to the detailed sound structure of the ambient language, as opposed to a structure of a non-native language (Kuhl et al., 1992; Polka & Werker, 1994). This indicated, according to the authors, that infants have a rather detailed representation of native segments. However, these studies were not aimed at studying word processing explicitly, so there is a possibility that infants do have detailed lexical representations of words, but they do not process words incrementally, i.e. they do not make use of the acoustic information before the offset of a word to the extent that adults do.

Modular or General Perceptual Restoration

There is evidence that the perceptual restoration phenomenon is not restricted to speech, to the modality of spoken language, or to the human species. Indeed, top-down
processing has been demonstrated in musical perception (DeWitt & Samuel, 2006), in the interpretation of American Sign Language (Schultz-Westre, 1985) as well as in the perception of starlings’ birdsong (Braaten & Leary, 1999).

Yet, due to psychoacoustic and methodological factors, some potential restrictions to the generality of the perceptual restoration phenomenon should be taken into consideration. As an example, for the perceptual restoration effect to take place, the class of the phoneme being masked and the nature of the masker (e.g. type of noise) as well as the amplitude of the masker have to be sufficiently similar (Newman, 2004). Originally a cough was used as a masker of the phoneme /s/ so that both the masker and the phoneme to be masked contained energy at several frequencies (Warren, 1970). Also the type of laboratory task facing subjects may differ in number of ways from e.g. everyday listening to speech in noisy environments. In the study by Newman (Newman, 2004), adults and school children were to detect interruptions presented at high rate (at slowest 2 per second). The interruptions also occurred at constant rate and were accordingly easily predictable by the subjects. Further, the stimuli in the experiment were presented over headphones and the adults were asked to type into a keyboard what they thought they heard, while the children were to repeat the sentences into a microphone. These recordings were later transcribed by an experimenter. In sum, the cognitive load intrinsic to the procedure of typing or repeating might have been rather demanding for the subjects. On contrary, an analysis of infants’ looking behaviour like in the studies by Fernald et al. (Fernald et al., 1998; Fernald et al., 2001) makes use of the infants’ spontaneous tendency to look at images of objects while hearing names corresponding to them.

Rationale for the Current Study

Although several studies have shown that adults and children possess the ability to identify words from partial phonetic information, less is known about the corresponding restoration ability in young infants. The aim of the present study was to investigate 11-16 months-old infants’ ability to identify spoken words on the basis of their initial sounds. The theoretical motivation for this study is inspired by analogue experiments performed on infant’s ability to predict other people’s action goals interpreted as a foundation of empathy and social understanding. The current study is further designed to investigate whether infants’ ability to predict phonetic events might be related to their age or to their productive vocabulary size. Therefore the youngest infants in the current study were chosen as representatives of subjects essentially lacking productive vocabularies.

As opposed to heavy-cognitive-load procedures, the current eye-tracking method (Tobii, http://www.tobii.com), just like the method used in the studies by Fernald et al. (Fernald et al., 1998; Fernald et al., 2001), takes advantage of the infants’ spontaneous tendency to look at images of objects while hearing names corresponding to them.

Method

The infants’ eye-movements were recorded as the subjects watched short video sequences displaying images of four familiar objects (a watch, a car, a ball and a teddy-bear). The objects were first introduced one at the time, along with a speaker voice presenting the Swedish names (target-nouns) of the objects embedded in carrier phrases. After these presentations, all the four objects were displayed simultaneously, one object per quadrant, while the speaker asked for one of the objects.

Subjects

A total of 17 infants participated in this study. Data from two infants were excluded due to interrupted recoding (failure to complete test session), resulting in 15 subjects (8 girls and 7 boys, age range 11.6-16.2 months, mean age 13.7 months). The infants were divided in two age groups, separated by the global median age – one group ranging from 11.6-13.5 months and the other from 13.5-16.2 months. All the subjects were primarily exposed to Swedish in their families. According to parental reports, none of the subjects had hearing abnormalities as revealed by BOEL-distraction test (“Blicken Orienterar Efter Ljud”) routinely used by Swedish Child health centres to screen all infants 7-10 months of age. The subjects’ receptive vocabularies were assumed to include the target-nouns used in the tests but nevertheless these target-nouns were explicitly presented in the video materials of this study. The infants were not expected to have these target-nouns in their productive vocabularies, although this may not be the case for some of the older infants.

Speech Materials

The target-nouns were selected from SECDI (Swedish Early Communicative Development Inventories) (Eriksson & Berglund, 1996) language assessment data base, which is a Swedish version of the MacArthur Communicative Development Inventories (CDI) based on parental reports. In order to assess age-appropriate test words, the “words and gestures” version of the database designed for children 8-16 months of age was used. In addition, only disyllabic target-nouns were selected and the words to be used in the test had to be matched regarding their initial phonemes. The target-nouns selected were:

- /klok:an/ (Watch)
- /nail:en/ (Teddy-bear)
- /bi:len/ (Car) (occurred only in manipulated version)
- /bol:en/ (Ball) (occurred only in manipulated version)

The disrupted target-nouns were prepared as follows: the first sound of /bi:len/ was extracted and concatenated with brown noise, the first sound of /bol:en/ was extracted and concatenated with brown noise, the first two sounds of /bil:en/ were extracted and concatenated with brown noise and the first two sounds of /bol:en/ were extracted and concatenated with brown noise resulting in:

- /b(i)/+NOISE (the /b/ carrying traces of /i/)
- /b(o)/+NOISE (the /b/ carrying traces of /o/)
- /bi/+NOISE
- /bo/+NOISE
Video Materials
The video materials (1 min 30 s) consisted of a BASELINE, an EXPOSURE, and a TEST phase, as follows (Figure 1):

- **BASELINE (20s):** The four objects were displayed to measure infants’ spontaneous preference before EXPOSURE. The audio played a lullaby to catch the infants’ attention towards the screen.

- **EXPOSURE (4 × 10s):** Each object was presented one at a time. A recorded female voice named the objects in carries phrases, uttered in Infant-Directed Speech style – “This is a X” or “Look at the X” (where X represents the actual target-noun).

- **TEST (6 × 5s):** Subsequently a split screen of the four objects was presented again. The voice asked for one of the objects – “Where is the X?” The target-nouns /klok:an/ (Watch) and /nal:en/ (Teddy-bear) were presented intact and repeated twice (TEST1 & TEST2). The other two target-nouns occurred only in their manipulated versions (TEST /bi:/ + NOISE, TEST /b(i)/ + NOISE, TEST /bo/ + NOISE, TEST /b(o)/ + NOISE).

There were four counterbalanced versions of the video materials in which the presentation order of the objects (during EXPOSURE) and questions (during TEST) was systematically varied to control for possible memory effects.

Procedure and Instrumentation
A Tobii 1750 eye-tracker integrated with a 17” TFT monitor was used in the measurements. For each subject the system was calibrated at the beginning of each session. The infant was seated facing the monitor at a distance of approximately 60 cm. The care-giver, listening to masking music through sealed head-phones with active noise reduction, sat by the infant, slightly behind, out of the infant’s visual field.

The equipment uses infrared light to generate gage measurements sampled at 50 Hz and its average nominal accuracy of gaze estimation is 0.5 degrees. The ClearView 2.2.0 software was used to store gaze data. The data collected was exported from ClearView to Matematica 5.2 and SPSS 14.0 for statistical analysis.

Results
The distribution of the looking times during the BASELINE showed a slight, non-significant, preference for Teddy-bear.

The data was assessed in terms of departures from the BASELINE to TEST expressed in looking times towards each object shown in the split screen. Figures 2-5 show detailed results for each of the target words pooled for all the infants. The results indicate that intact target nouns were recognized in a much more accurate way than the partially masked words. Significant departures/tendencies according to paired samples test (2-tailed) from BASELINE to TEST are shown in Table 1.

Additional phonetic information (i.e. partially masked words represented by their first two sounds as opposed to by one sound only) did not seem to improve the infant’s ability to predict the target word.
Figure 3: **Target Teddy-bear.** CI (95%) from left to right indicate looking time (ms) during BASELINE, TEST 1 (1st rep. of /nal:en/), and TEST 2 (2nd rep. of /nal:en/).

Figure 4: **Target Car.** CI (95%) from left to right indicate looking time (ms) during BASELINE, TEST (/bi/+NOISE), and TEST (/b(i)/+NOISE).

Figure 5: **Target Ball.** CI (95%) from left to right indicate looking time (ms) during BASELINE, TEST (/bo/+NOISE), and TEST (/b(o)/+NOISE).

The two age groups (11.6 to 13.5 and 13.5 to 16.2 months) behaved in about the same way, although there was a slight advantage for the older group in responding to intact words.

**Summary and Discussion**

Infants seem to possess prediction skills for anticipation of other people’s goal-directed behaviour already by the age of 12 months (Falck-Ytter et al., 2006). The mirror-neuron system (MNS) coding for learning to do an action from seeing it done, and functioning as the neurological correlate of these and similar results, is suggested to exist both in monkeys (area F5) (Rizzolatti & Arbib, 1998) and humans (Broca’s area) (Fadiga et al., 1995). As an extension, the MNS hypothesis of social cognition (Rizzolatti & Craighero, 2004) proposes this system to constitute a neurological base for coding social understanding and language abilities. Evolutionarily, the coding of hand-actions is progressively assumed to be transferred to mouth-actions in general and further through ingestive behaviour (such as breaking food) to communicative mouth-actions (such as lip smacking) in particular. Accordingly, the original function of *audio-visual* mirror neurons might have transferred to respond to sound only, leading to a separate *auditory* mirror system on the top of the original audio-visual one (Rizzolatti & Craighero, 2004).

Experimental evidence points at commonalities between the development of prediction skills of action goals for motor and speech events. Infants, as young as 18-21 months-old, seem to be able to predict partial acoustic information (Fernald et al., 2001). In addition, these results suggested that recognition of non-complete acoustic speech signals is related to infants’ lexical development (productive vocabulary sizes) and/or to their age (Fernald et al., 1998). Further, the “perceptual restoration” phenomena – which comprises that the observer perceives the word/song/signing to continue behind the noise and restore what ever is

<table>
<thead>
<tr>
<th>Intact target nouns</th>
<th>Interrupted target nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEST 1 /klok:an/ (Watch)</td>
<td>t(13)=2.482, p&lt;0.027</td>
</tr>
<tr>
<td>TEST 2 /klok:an/ (Watch)</td>
<td>t(12)=2.102, p&lt;0.057</td>
</tr>
<tr>
<td>TEST 1 /nal:en/ (Teddy-bear)</td>
<td>t(14)=2.057, p&lt;0.059</td>
</tr>
<tr>
<td>TEST 2 /nal:en/ (Teddy-bear)</td>
<td>t(12)=0.974, p&gt;0.349</td>
</tr>
</tbody>
</table>
missing – is suggested to function in a domain- and species-
general manner (e.g. to exist also in music, sign-language, and in birdsong).

The current study was set up to investigate young infants’
(age range 11-16 months) ability to use partial phonetic
information. The choice of age-continuum and the youngest
infants’ supposedly non-existing productive vocabularies
were a result of the endeavour to bring light on the
discussion of whether infants’ ability to predict acoustic
information is correlated to productive vocabulary size
and/or age. Caution has been taken concerning potential
methodological pitfalls regarding the class of phonemes to
be masked and the energy levels of the masker at different
frequencies, as well as regarding the type of laboratory task
facing the infants. The brown noise used (containing energy
at all the frequencies) was simply added – as opposed to
superimposed on the phonemes to be masked – to the first
phoneme/phonemes after excising the last phonemes of the
signal. In this way the masker was prepared to function
effectively assuring the perceptual restoration effect to be
able to take place. In addition, the interrupted signals were
not presented in a predictable or in a cognitively demanding
manner. The infants’ spontaneous tendency to look at
objects while hearing names referring to them was used.

To test the perceptual restoration effect’s modularity to
speech processing versus it’s generality as a part of auditory
behaviour, studies on infant’s restoration of non-linguistic
stimuli, as well as studies on non-human mammals’
restoration of linguistic stimuli, are to be performed in
future experiments.

Acknowledgments

Research supported by The Swedish Research Council, The
Bank of Sweden Tercentenary Foundation, European
Commission, and Birgit and Gad Rausing’s Foundation.
The authors are grateful to a group of speech-therapist
students (Logopedprogrammet 2005) for performing the
subject running.

References

of missing birdsong segments in European starlings.
Psychological Science, 10, 162-166.
knowledge-based expectations in music perception:
Evidence from musical restoration. Journal of
Experimental Psychology, 119, 123-144.
Communicative Development Inventory -- words and
gestures. First Language, 19, 55-90.
Motor Facilitation During Action Observation - A
Magnetic Stimulation Study. Journal of Neurophysiology,
73, 2608-2611.
Infants predict other people's action goals. Nature
Neurosci., 9, 878-879.
Fernald, A., Pinto, J. P., Swingley, D., Wineberg, A., &
processing by infants in second year. Psychological
Science, 9, 228-231.
a word is enough: infants can recognize spoken words
using partial phonetic information. Child Dev., 72, 1003-
1015.
Ferrari, P. F., Gallese, V., Rizzolatti, G., & Fogassi, L.
(2003). Mirror neurons responding to the observation of
ingestive and communicative mouth actions in the
monkey ventral premotor cortex. European Journal of
Neuroscience, 17, 1703-1714.
Grafton, S. T., Arbib, M. A., Fadiga, L., & Rizzolatti, G.
(1996). Localization of grasp representations in humans
by positron emission tomography .2. Observation
compared with imagination. Experimental Brain
Research, 112, 103-111.
Keysers, C., Kohler, E., Umilta, M. A., Nanetti, L., Fogassi,
L., & Gallese, V. (2003). Audiovisual mirror neurons and
action recognition. Experimental Brain Research, 4, 628-
636.
Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese,
V., & Rizzolatti, G. (2002). Hearing sounds,
understanding actions: action representation in mirror
Kuhl, P., Williams, K., Lacerda, F., Stevens, K. N., &
phonetic perception in infants by 6 months of age.
Science, 255, 606-608.
MacNeilage, P. & Davis, B. L. (2000). Deriving speech
from nonspeech: a view from ontogeny. Phonetica, 57(2-
4), 284-296.
Polka, L. & Werker, J. F. (1994). Developmental changes in
perception of nonnative vowel contrasts.
restoration: Sign restoration in American Sign Language.
Milwaukee, WI: University of Wisconsin-Milwaukee.
and phonemic restorations. Perception and
Psychophysics, 9, 358-362.
and confusions. Scientific American, 223, 30-36.
Emerging Linguistic Functions in Early Infancy

Francisco Lacerda*, Eeva Klintfors, Lisa Gustavsson
Ellen Marklund and Ulla Sundberg
Department of Linguistics
Stockholm University
SE-106 91 Stockholm
Sweden
* frasse@ling.su.se

Abstract
This paper presents results from experimental studies on early language acquisition in infants and attempts to interpret the experimental results within the framework of the Ecological Theory of Language Acquisition (ETLA) recently proposed by (Lacerda et al., 2004a). From this perspective, the infant’s first steps in the acquisition of the ambient language are seen as a consequence of the infant’s general capacity to represent sensory input and the infant’s interaction with other actors in its immediate ecological environment. On the basis of available experimental evidence, it will be argued that ETLA offers a productive alternative to traditional descriptive views of the language acquisition process by presenting an operative model of how early linguistic function may emerge through interaction.

1. Introduction

Previous studies of the young infant’s ability to learn names of objects presented under controlled naturalistic settings have demonstrated that by about 7 to 8 months of age, infants are capable of interpreting arbitrary words as names of visual objects, provided the words and the objects co-occur consistently. For instance, a study by Gogate and Bahrick (Gogate & Bahrick, 1998) indicates that 7 month-old infants are able to explore audio-visual co-occurrences to establish arbitrary word-like associations between isolated speech sounds and objects. In addition, a more general assessment of the impact of audio-visual synchrony (Prince et al., 2004) strongly suggests that synchronic events may expose linguistically relevant audio-visual relationships. But while the young infant’s ability to establish sound-object links offers god support to the notion that association processes are likely to underlie early language acquisition, accounting for the language acquisition process in terms of relatively simple associative processes involving isolated words is problematic because it may lack general ecological relevance. Indeed, as often pointed out by scientists criticizing the emergentist views of the language acquisition process (Lidz, Gleitman, & Glei tman, 2003), words representing the names of objects available to the young infant tend to be embedded in utterances rather than uttered in isolation, an aspect that necessarily reduces the ecological relevance of experimental studies reporting referential learning from words presented in isolation. Thus, to further investigate the extent to which general association processes might underlie early language acquisition in ecologically relevant adult-infant interaction settings, a series of experiments were set up in which the target words were integrated in natural sentences (as those typically heard by infants) and arbitrarily combined with visual objects simultaneously accessible to the infants.

The present paper will argue that early language acquisition can indeed be seen as the result of an interactive process between the infant and its environment, through which the infant picks up linguistic regularities afforded in the ambient language. In the following we try to provide an empirical basis for our emergentist views of the early language acquisition process by reviewing some of our experimental studies addressing different aspects of early language acquisition in infants and examining the characteristics of the infant-directed. We will first review an experiment designed to test how different linguistic factors may influence the infant’s ability to derive word-object relationships from exposure to naturalistic audio-visual contingencies. Thereafter we will examine the characteristics of infant-directed speech from the perspective of the ETLA. Finally, we will address the issue of necessity of general-purpose versus language-specific processes underlying the infant’s ability to link visual and auditory information and form productive linguistic representations.

2. Emerging word-object associations

To investigate the generality of the word-object association process, a series of experiments were carried out to investigate the infant’s ability to derive the names of objects from experience with audio-visual stimuli, where natural sentences conveying implicit referential information are presented simultaneously with the visual images of the objects they refer to. One setup of these studies was already described in (Gustavsson et al., 2004) and will be briefly reviewed here.

This study used a Visual Preference procedure similar to the procedures used by Fernald and her colleagues (Fernald, Swingley, & Pinto, 2001; Swingley, Pinto, & Fernald, 1999). In general terms, the procedure can be described as inducing the infant’s response from its looking time towards alternative pictures displayed simultaneously and where one of the pictures is associated with the expected response.
2.1 Speech materials

The speech materials were Swedish sentences recorded by a female native speaker of Swedish. The utterances introduced non-words as names of the objects being displayed on the screen. Nine films were created to include all the possible combinations of position of the target word (initial, medial or final position in the utterance) and the utterances focal accent (falling on the utterances initial, medial or final words). The syntactic structure of the utterances was different from film to film but within each film the position of the target word and the part of the utterance receiving focal accent was kept constant. Furthermore, although the utterances within each film were structurally equal, the non-target words were different from utterance to utterance in an attempt to mimic the variation typically observed in natural utterances. Examples of the utterances presented in two of the nine films are shown in table 1, where the focal accent is indicated by boldface and the position of the target word by XXX. For the placement of the target word and focal accent, the utterances were divided in three regions – initial, medial and final. The initial and final positions were defined by the first and the last word in the utterance. The medial position was defined as the remaining part of the utterance.

<table>
<thead>
<tr>
<th>Film 1</th>
<th>Film 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>target word: final</td>
<td>target word: final</td>
</tr>
<tr>
<td>focal accent: final</td>
<td>focal accent: medial</td>
</tr>
<tr>
<td>Titta här är sötta XXX</td>
<td>Det är den söta XXX</td>
</tr>
<tr>
<td>Se på den lilla XXX</td>
<td>Se på den lilla XXX</td>
</tr>
<tr>
<td>Titta på fina XXX</td>
<td>Titta på fina glada XXX</td>
</tr>
<tr>
<td>Kolla in den glada XXX</td>
<td>Kolla glada XXX</td>
</tr>
</tbody>
</table>

Table 1. Example of the Swedish utterances presenting the target words. The target word is represented by XXX, standing for the non-words “Kucka” and “Dappa”. Focal accent is represented by boldface.

Each of the nine films was organized in three phases – baseline, exposure and test.

In the baseline phase, still images of two puppets were displayed side by side in a split-screen. The duration of the baseline phase was 30 s. During the baseline phase an especially composed short instrumental lullaby (Anna Ericsson, 2004) was played, starting approximately 2 s after the onset of the visual display and finishing about 2 s before the end of the baseline phase. The infant’s looking towards each of the puppets during this phase was used as a measure of the subject’s preferential bias towards the puppets.

During the exposure phase, two short 20 s video sequences were played to show each of the puppets per se, introduced by the sentences referring to the particular puppet being displayed (see table 1). The sentences were evenly distributed throughout the duration of each video sequence. The first sentence started about 1 second after the onset of the visual display and the last sentence finished about 1 s before switching to the next video sequence. These video sequences were presented after each other, switching from one puppet to the other. The total duration of the exposure phase was 120 s, during which each of the individual video sequences was presented 3 times. The infants’ looking time towards the each of the puppets was taken as a measure of attention during the exposure phase.

In the test phase the two puppets were again displayed in a split-screen similar to that of the baseline but now the audio track played questions like “Where is XXX?” or “Can you see XXX?”, where XXX was the name of one of the puppets, implicitly introduced in the descriptions presented during the exposure phase. The test phase was 30 s long, just as the baseline phase.

2.2 Subjects

A total of 49 infants participated in the study. Some of the infants participated in more than one session, adding up to 78 sessions. The results presented here come from a total of 75 sessions, distributed as indicated in the table below. The ages of the subjects at the time of their participation in the sessions ranged from 201 to 278 days (mean age was 239 days, s.d.=15 days). The age distribution for this sample was nearly Gaussian (skewness=0.180; kurtosis=0.503).

<table>
<thead>
<tr>
<th>Target word position</th>
<th>Focal accent</th>
<th>initial</th>
<th>medial</th>
<th>final</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>medial</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>final</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>22</td>
<td>25</td>
<td>75</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Number of data points used for specific combinations of target word and focal accent positions.

The selection criterion for the sessions above required that the looking bias towards one of the puppets did not exceed 35% of the baseline time.

2.3 Procedure

The subjects were video recorded during the experiments, using a camera placed just above the display they were looking at. To register the actual images that the infant was looking at and give the possibility of re-analysis, the film being displayed was mixed onto the upper left corner of the image of the subject’s face. This overlapping image used about 1/16 of the screen area and did not interfere with the image of the face of the infant. A time stamp placed at 40 ms intervals was also recorded on the lower right corner of the screen. This time stamp was subsequently used to compute a session-relative time, allowing the line up the start of the video films from different subjects.

In this experiment, the looking times towards each of the puppets were measured manually, frame by frame, a very time consuming procedure. The separation between the target images was about 30", which was enough to allow clear decisions on which side of the screen the subject was momentarily looking at. Three levels of looking were coded – left, right and off.

On the basis of these codes, a “pre-to-post exposure gain” variable was defined as the net increment in looking time towards the puppet used as target.
Gain = Tgt − TgtB

where Tgt is the total looking time towards target puppet and TgtB is the total looking time during baseline towards the puppet that would become the target in the test phase.

3. Results

The results from the first sessions in which the 36 selected subjects participated are shown in figure 1, grouped according to the placement of the focal accent and the position of the target word in the utterances.

Given the reduced number of subjects each condition and the typical variance observed in this type of experiments, it is perhaps not surprising that no significant main effects for the target word or placement of the focal accent could be observed. As stated above, this analysis was carried out on a selection of all the sessions in which the infant’s initial bias towards any of the puppets was less than 35% of the total baseline time. Further analyses using all the available data from the 78 sessions did not change appreciably the pattern displayed on figure 1. The main difference was a broadening of the confidence interval for medial target word position with medial focus, due to an extreme negative gain outlier resulting from a strong bias towards the puppet that would function as target. A non-parametric display of the same data is shown in figure 2. The dependence of the median values on the target word and focal accent position is in good agreement with the pattern displayed in figure 1. There were no significant main effects or interactions for word position and placement of the focal accent. However a tendency for longer looking times was observed for the target word in focal position ($F(1,73)=2.957$, $p<0.090$). If the case of the target word in final position, with a focal accent in the initial position of the sentence is excluded, then a significant effect of the placement of the target word in focus is obtained ($F(1,65)=4.075$, $p<0.048$). Furthermore, there was a significant difference between the mean looking times for the group of sentences with the target word in focal position and sentences with the target word in final position and focal accent in initial position ($F(1,73)=5.579$, $p<0.021$).

4. Discussion

While there were no overall significant differences when considering all the data at once, the response pattern displayed in figures 1 and 2 strongly suggested that target words in focal position might have been easier to associate with the corresponding puppets than when focal accent did not fall on the target word. This means that 8-month-old infants seem to be able to pick up relevant linguistic information by listening to the word that is placed in focal position. An unexpected effect was however observed when the target word was in non-focal final position but the utterance had initial focal accent. It appears that the initial focal accent may have primed the infants to attend to the utterance, prompting the subjects to retrieve the less prominent target word delivered in sentence final position.

In summary, the results of this experiment seem to indicate a general ability to link recurrent target words with visual objects that are simultaneously available to 8-month-old infants, providing the ground for the linguistically relevant referential function. The fact that the strength of the responses varied significantly for different combinations of focal accent target word placement further suggests that the infants’ ability to pick up the linguistic referential function was modulated by prosodic patterns and primarily contingent on the coherence in the placement of the focal accent and the target word. An implication of this is that by 8 months of
The emergence of the linguistic referential function suggested by the study reported in the previous section may be seen as a consequence of a general multi-sensory representation process through which synchronic multi-sensory information is spontaneously associated, thereby exposing implicit cross-modal regularities (Lacerda et al., 2004a; Lacerda et al., 2004b; Lacerda, 2003). Because the efficient use of spoken language is based on the ability to relate sound symbols (however variable) to objects perceived (primarily but not exclusively) by other senses, a systematic (or at least predictable) link between the sound code and the objects it refers to must exist at some level of representation (Minsky, 1985). Note that in line with ETLa (Lacerda et al., 2004a), such a sound code is a generic reference to the concrete auditory impression of a word or a lexical phrase as a whole, not to the word’s representation in terms of linguistic concepts like phonemes or syllables nor to the sequence of words that may build up the lexical phrase. In this perspective, words, syllables and phonemes are an emergent consequence of the combinatorial pressure imposed by increasing representation needs (Nowak, Plotkin, & Jansen, 2000; Lacerda, 2003).

To address the issue of the generality of cross-modal links in infancy, we carried out a study to investigate the infant’s ability to use temporal synchrony to relate ecologically relevant auditory and visual speech information, the infant’s ability to relate ecologically relevant synchronic non-speech audio and visual information and the infant’s ability to relate synchronic non-speech audio with speech (articulatory) visual information.

The background for the present experiments is an early study by Kuhl and Meltzoff’s study (Kuhl & Meltzoff, 1982) showing that 18 to 20 weeks-old infants can pick up the correlation between acoustic and articulatory characteristics of speech sounds. In their study the infants were exposed to a split-screen displaying two faces, one articulating /a/ and the other articulating /i/, while an audio signal consisting of either one of those vowels was played. Their results indicated significantly longer looking times towards the face whose articulation was consistent with the audio signal.

Also (Bahrick, 2004) carried out a study in which 5 months-old infants were tested on their ability to discriminate between different phenomena involving changes in rhythm or tempo. The tests were organized in three situations: (1) a multimodal situation, where a plastic hammer was seen while the sound of the hammer hitting a
surface was heard, (2) a unimodal situation where only the sound of the hammer was heard, and (3) a unimodal situation where the hammer was only seen but not heard. The rhythm of the events was subsequently manipulated in each of these three situations in order to create novel situations that the infants might discriminate. The results indicated that only the 5 month-old infants who received the bimodal redundant stimulation could detect the rhythm changes. According to other studies by Bahrick, infants tend to be less dependent on redundant information the older they get.

Our study attempted to expand the findings of Kuhl and Meltzoff (Kuhl et al., 1982) and Bahrick’s by investigating the ability of 6 to 8 months-old Swedish infants to perceive synchronous visual and auditory input, for both speech and non-speech events. We also introduced a methodological improvement by using a high resolution eye-tracking system, with a maximum resolution of about 0.5°, which allowed the presentation of four images on a single screen during the test phase instead of the two alternatives used by Kuhl and Meltzoff, thereby reducing to 25% the spontaneous chance level of looking at one of the images. Just as in Kuhl and Meltzoff’s case, we hypothesized that the infants would look significantly longer towards the images displaying motor activities coherent with the heard speech or non-speech signals.

6. Method

After a short calibration of the eye-tracking system, the infants were exposed to a short video film while their eye-movements were registered throughout the session. For this paper, only the infants’ average looking times towards the different quadrants of the split-screen will be considered for statistical analysis. However, the eye-tracking data was collected with high enough temporal and spatial resolution to allow a detailed study of the infants’ visual strategies but those results will be reported in a future paper.

6.1 Subjects

Of the forty infants who participated in this study four had to be excluded due to calibration errors. The resulting in 36 subjects (13 boys and 13 girls) aged 25-33 weeks (mean age 28.5 weeks). The subjects were randomly selected from the National Swedish address database (SPAR) targeting 6 to 8 months-old infants whose parents lived in the Stockholm metropolitan area.

6.2 Stimuli

The infants were exposed to a film showing a female actress against a blue background. The film consisted of four sequences: (1) a baseline for speech articulations, (2) a test phase for audio-visual coherence in speech stimuli, (3) a baseline for non-speech gestures and (4) a test phase for audio-visual coherence in non-speech stimuli.

In the speech part or the experiment the baseline consisted of four identical still images on a split-screen showing the actress’s face. The baseline of the non-speech part of the experiment was an animated video sequence showing four different tempos of hand clapping, one in each quadrant. This baseline sequence was identical to the one to be used in the test phase, but with a silent sound track.

In the test phase for the speech stimuli (figure 4), the actress was again shown on a four quadrant split-screen articulating the vowels [a] and [y] and the syllables [ba] and [by]. For periods of 20 seconds, the speech signal was synchronized with the film shown in one of the quadrants.

In the first 20 seconds the speech signal consisted of repetitions of the syllable [by] and the organization of the four video tracks was (target position in boldface)

```
ba
by
y
a
```

Figure 4. Example of the speech sound part of the experiment. The actress is articulating [ba] (UL), [by] (UR), [y] (LL) and [a] (LR). The audio played was the syllable [by], i.e. the target image was UR.

Directly after the vowel [a] was presented in the next 20 seconds and the position of the visual targets was

```
by
ba

a
y
```

After this the [by] syllable was repeated as target but the visual target was placed in another quadrant

```
y
a

ba
by
```

Finally the [a] was presented again and the visual target once more relocated

```
a

y

by
ba
```

The utterances were produced with rise-fall f0 contours and the target images were placed in different quadrants for each of the four 20 seconds sequences, as shown in the tables above.

For the test phase with non-speech stimuli (figure 5), the actress was shown on a split-screen clapping hands in different tempos. The tempos were 157%, 101%, 63% and 49% of the original recording tempo. The audio was manipulated to 101% of the original recording tempo, thus synchronized with one of the images shown.

The location of the videos on the split screen is given in the tables below.
Baseline for rhythm (clapping movements with no sound)

<table>
<thead>
<tr>
<th>101%</th>
<th>49%</th>
</tr>
</thead>
<tbody>
<tr>
<td>63%</td>
<td>157%</td>
</tr>
</tbody>
</table>

Test phase for clapping rhythm. The spatial organization of the videos is the same as during baseline except that now the sound track corresponding to the 101% speed plays the clapping sounds.

Baseline for coupling of clapping sounds with visual displays of [by] utterances. The display shows four images of the actress uttering [by] at different speeds. During the baseline the soundtrack is silent.

<table>
<thead>
<tr>
<th>63%</th>
<th>49%</th>
</tr>
</thead>
<tbody>
<tr>
<td>101%</td>
<td>157%</td>
</tr>
</tbody>
</table>

The video films for this test phase were identical to those of the baseline but now the sound track played a clapping sound synchronized with the articulatory movements shown on the lower left quadrant.

Half of the subjects were exposed to the speech sound part of the experiment first, followed by the hand-clapping part. The rest of the subjects were exposed to the two parts in reversed order.

6.3 Material

The equipment used for tracking the infant’s eye movements was Tobii 1750 eye-tracker integrated with a 17” TFT monitor. The system uses low intensity infrared light to create a static reference frame on the spherical surface of the eye and derives a gaze vector from the relative position of the pupil within that frame. The system performs gaze measurements 50 times per second and with a nominal accuracy of 0.5°. The eye-tracking data generated by the ClearView 2.2.0 software package that comes with the system was subsequently analyzed using Matematica 5.1 and SPSS 13.0.

6.4 Procedure

The experiments were carried out in a dimly lit studio where most of the light came from the screen connected to the eye-tracking system. The brightness of the display on this screen was enough to draw the infant’s attention towards the stimuli being presented. The infant sat in front of the screen at a distance of approximately 60 cm. The parent sat in the studio slightly behind and outside the infant’s visual field and listened to music played through sound-isolating head-phones equipped with active noise reduction. Before recording the gaze the system was calibrated using the infant’s fixations on special purpose calibration points that were displayed on an otherwise empty screen. The calibration procedure was typically carried out in less than one minute.

7. Results

An example of the infants’ responses during one of the speech sound conditions is displayed in figure 6. Each panel corresponds to a quadrant on the test screen. The condition illustrated in figure 6 refers to the 20 s video sequence during which the infants heard the vowel [a] for the first time in the session. The video corresponding to the sound track was displayed on the lower left corner (LL) of the screen. The curves in each of the panels show the percentage of infants who, at a given time throughout the 20 s of that test phase, were looking at the quadrant represented by the panel. The curves indicate a looking preference towards the upper quadrants, with a slight dominance for the upper right quadrant, displaying the articulation of [ba] syllables. The upper left quadrant, receiving the next highest percentage of looking time through this test phase, displayed the articulation of [by]. The correct visual target was in fact displayed on the lower left quadrant in this test phase and appears to have in fact received the lowest average percentage of looking time.

Figure 5. Hand-clapping in four different tempos: 101% (UL), 49% (UR), 63% (LL), and 157% (LR) of the original tempo. The sound of hands clapping was synchronized with the target image (UL).

Figure 6. Percentage of infants looking towards each of the quadrants on the screen as a function of time. A running time-window of 200 ms was used. First presentation of [a]. The target image was placed on the lower left (LL) quadrant in this case.
To obtain the net individual gains in looking time towards each of the quadrants shown during the test phases, a repeated measures analysis of variance was performed using the looking times towards a given quadrant during the baseline and the test phase. The results, using film order as between subject’s order, indicated a significant gain for the upper left quadrant displaying [ba], when [by] was heard ($F(1,34)=5.243, p<0.028$). There was no significant interaction with film order but film order was a significant between subjects effect ($F(1,34)=4.303, p<0.046$). These results suggest thus that the infants matched the visual image of [ba] with the sound of [by], although the group of infants who started the session seeing the clapping sequences performed not as well as the group seeing first the speech stimuli. Another significant gain in looking behaviour was observed for seeing [ba] and when listening to [a] ($F(1,34)=6.196, p<0.018$). No significant interaction with film order or significant effect of film order was observed in this case.

The analysis of the looking behaviour during the baseline phases indicates that the infants tended to look most of the time towards the upper quadrants. To compensate for this bias, the results from were sorted in terms of video materials being displayed, rather than the quadrants on which they appeared. Thus, the total looking time towards visual [ba] while listening to [by], for instance, was computed by adding the gain in looking time towards the upper left quadrant during the first test phase where [by] was heard, with the gain in looking time towards the lower left quadrant during the other test phase during which [by] was played. The results from this type of analysis are shown in figures 7 and 8. ANOVA models using the individual subject’s looking times towards each of the visual displays shown in figure 7 revealed a within-subjects significant linear trend in looking behaviour towards [ba], [by], [a] and [y] ($F(1,35)=7.235, p<0.011$). A very significant within-subjects linear trend was also found for the pattern displayed in figure 8 ($F(1,35)=27.507, p<0.0005$).

7.1 Non-speech sounds

The same type of analysis was carried out for the video films involving clapping sounds. In this case all the four quadrants displayed the same type of action but the action was performed with different repetition rates. The results from the infant’s matching between clapping sounds and the videos showing the actress clapping at different rates are shown in figure 9. The gain in looking time is greatest for the upper left quadrant, which also is the quadrant showing the clapping movements in synchrony with the sound.

In the other situation involving clapping sounds the videos displayed the actress rhythmically uttering [by]. The results in this case did not show maximum looking time gain towards the lower left quadrant containing the utterances synchronized with the clapping sounds, as illustrated in figure 10. Instead, the maximum looking time was towards the lower right quadrant. However,
when the looking behaviour is organized as a function of
the repetition tempo, a pattern of increasing looking times
for increasing frequency in the repetition of the
articulatory movements emerges. This is significant linear
trend ($F(1,35)=9.365, p<0.004$).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure10.png}
\caption{Average looking time gains (in ms) and 95\% confidence intervals for looking towards [by] while hearing clapping sounds.}
\end{figure}

8. Discussion

The results of these experiments do not support the strong
notion that 8-month-old infants might be able to establish
phonetically relevant correspondences between speech
sounds and their underlying articulatory movements. Indeed, rather than looking at the quadrants displaying the
articulatory movements associated with the speech sounds,
the infants’ preferences seemed to follow the salience of
the articulatory displays on the quadrants (ba->by->a->y).
But this was not because they were unable to detect
synchrony in general terms. As demonstrated by the tests
with clapping sounds, the infants were able to pick up the
correct audio-visual synchrony when clap sounds and
images were present but they appear to treat speech
sounds (or the articulatory movements associated with
speech sounds) in a different way than non-speech sounds.
In fact, the infants failed to detect synchrony between the
non-speech sound and the synchronic articulatory
movements. They looked instead longer towards the video
film showing the most rapid alternations between closed
and open lips, a response that is in line with the results
from the “speech sound part” of the experiment.

9. Conclusion

Taken together, the two experiments reported here suggest
that 8 month-old infants may be using unspecific
associative functions to pick up relevant linguistic
information on the basis of multi-sensory regularities
available in their immediate linguistic environment. If this
is true, the infant’s success in acquiring the relevant
linguistic functions is in line with ETLA and may be more
dependent on the structure of its linguistic ambient than on
the unfolding of a language acquisition program. In

addition, from the point of view of epigenetic robotics this
may be a general productive approach, worth to pursue
(Dominey & Boucher, 2004). Indeed, given the repetitive
characteristics of speech directed to infants (IDS) about 3
months of age, ETLA suggests that it may be possible to
derive meaning from the recurrent co-occurrences of
auditory and other sensory information representing the
infant’s immediate linguistic environment. In our MILLE-
project, we are currently making efforts to model the early
stages of language acquisition exploring the acoustic
regularities available in repetitive IDS.

Acknowledgements

Research carried out with grants from the Swedish
Research Council and the Bank of Sweden Tercentenary
Foundation.

References

Bahrick, L. E. (2004). The development of perception in a multimodal
environment. In G. Breckenridge & A. Slater (Eds.), Theories of Infant
Development (pp. 90-120). New York: Blackwell Publishing.

perception and language acquisition in a perceptually grounded
robot. In L. Berthouze, H. Kozima, C. Prince, G. Sandini, G.
Stojanov, G. Metta, & C. Balkenius (Eds.), (pp. 43-50). Genova,
Italy.

Fernald, A., Swingley, D., & Pinto, J. P. (2001). When half a word is
enough: infants can recognize spoken words using partial phonetic
information. Child Dev., 72, 1003-1015.

facilitates learning of arbitrary relations between vowel sounds and
objects in seven-month-old infants. J.Exp.Child Psychol., 69, 133-
149.

Gustavsson, L., Sandberg, U., Klintfors, E., Marklund, E., Lagerkvist, L.,
& La (2004). Integration of audi-visual information in 8-month-old
infants. In L. Berthouze, H. Kozima, C. Prince, G. Sandini, G.
Stojanov, G. Metta, & C. Balkenius (Eds.), (pp. 143-144). Genova,
Italy.


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

Lacerda, F., Klintfors, E., Gustavsson, L., Lagerkvist, E., Marklund, E.,

Lacerda, F., Marklund, E., Lagerkvist, E., Gustavsson, L., Klintfors, E.,
& Sundberg, U. (2004b). On the linguistic implications of context-
bound adult-infant interactions. In Genova: Epirob 204.

Lacerda, F. & Sundberg, U. (1996). Linguistic strategies in the first 6-
months of life. In (pp. 2574). JASA.

matters: verb learning and the footprint of universal grammar.
Cognition, 87, 151-178.

Regis (Ed.), Extraterrestrials: Science and Alien Intelligence.
(Cambridge, MA: Cambridge University Press.


Prince, C., Hollich, G., Heldcr, N., Mislove, E., Reedy, A., Saluske, S.,
et al. (2004). Taking Synchrony Seriously: A Perceptual-Level
Model of Infant Synchrony Detection. In L. Berthouze, H. Kozima,
C. Prince, G. Sandini, G. Stojanov, G. Metta, & C. Balkenius (Eds.),
(pp. 89-96). Genova, Italy.