Non-auditory cognitive capabilities in computational modeling of early language acquisition

Okko Räsänen

Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

okko.rasanen@aalto.fi

Abstract

Computational models of early language acquisition (LA) play an important role in understanding the acquisition and processing of spoken language. Since language is an extremely complex phenomenon, computational studies typically address only a specific aspect of the LA at a time. This calls for a huge number of assumptions regarding the other cognitive processes of the learning system, and these assumptions can have significant consequences to the ecological plausibility of the simulations. In this paper, we review the developmental status of a number of cognitive processes during the first year of infant's life that are typically involved in the computational simulations of LA. How these findings are related to the plausibility of different simplifications and assumptions in computational models are also discussed.

Index Terms: language acquisition, computational modeling, statistical learning

1. Introduction

Computational models of language acquisition (LA) are needed in order to truly understand the principles behind processing of spoken language (see [1] for a review). Not only can they provide support for, or falsification of, existing theories, but they also help to formulate new hypotheses about how LA process may take place. One of the central topics in the recent LA research has been *distributional learning*, i.e., the finding that a great deal of infant language learning can be explained by general statistical learning mechanisms without a need for innate language specific processes or biases. Naturally, computational models of LA play a central role in the study of this field due to the advocated metaphor of human mind as a computational system, and due to the fact that the basic principles and processes of statistical learning are analogous to those of machine learning and unsupervised pattern discovery.

However, a general problem with computational modeling of LA is that language is an extremely complex phenomenon including physical, (neuro)physiological-, cognitive-, behavioral- and cultural factors that all shape the outcome of human spoken languages. On the other hand, simulations typically attempt to address only a specific aspect of the language learning process. Isolating one part of the entire process for closer study means that strong assumptions have to made regarding the excluded aspects of the process, either by prescribing how the other aspects behave, or at least assuming that they will not have significant consequences on the results of the current study. Compiling findings from numerous isolated studies also requires that the assumptions of these studies are mutually compatible.

If one wishes to work towards integrative models of LA and to study the extent of the role of distributional learning, acknowledging the ecological plausibility of different simplifications and assumptions used in the computational experiments is necessary. Following the definition in [2], these assumptions mainly concern non-linguistic aspects of cognition that can be considered as *language-experience-independent mechanisms*. Their development is not usually addressed in the models focused on LA, but are assumed to be innate or previously learned by the agent. Justifications for the assumptions are rarely given, although this is not surprising given the already large scope of theoretical and computational considerations required in these computational studies.

In order to provide understanding on the plausibility of the existing computational models and to encourage ecologically plausible experimental settings also in the future studies, the current work reviews the concurrent developmental states of a number of cognitive processes that are often required in computational studies of early LA. These processes and the related assumptions include 1) the ability to perceive emotional feedback as a reward in reinforcement learning, 2) visual processing in the perception of the objects and events that are being discussed, 3) the ability of the learner to follow the attention of the caregiver, and 4) the ability to compute statistics over multiple interaction situations in order to learn the mapping between the auditory word forms and their referents [1]. The four topics are discussed in their respective order. Finally, some conclusions are drawn.

2. Ability to perceive emotional feedback

The use of reinforcement learning techniques in LA studies (e.g., in caregiver-learner interaction) is supported by the findings that the emotional states of others are perceivable to the infants at four months of age through facial expressions [3] and tone of speech [4]. The preference for rhythm and ability to perceive emotions from vocal gestures are an aspect of language that may not necessarily be innate but can be actually considered as a special case of supervised learning: the preference for rhythm and positive emotional tone may be learned prenatally. It is known that prenatal human infants are able to hear sounds in the womb [5], including low-passed versions of external sounds (e.g., rhythm of speech) and sounds originating from the mother. Simultaneously, the embryo is coupled to its mother's hormonal system via the placenta, enabling associative learning between auditory patterns and emotional states of the mother. For example, positive communicative situations may induce emotionally colored associations to a speaking style specific for positive affective state. In [4] it was proposed that the grounding through physiological coupling would explain why the newborn

infant is already able to differentiate speech with positive or negative emotional tone, where the emotion specific intonation pattern may depend on the native language [6] (see also [7]). The prenatal grounding to physiological states may also explain why infants enjoy rhythmic patterns that are typical to their native culture [8]. Also, the emotional signaling may be important in the interaction between the caregiver and the infant, and interaction styles have been shown to correlate with language learning [9]. Interaction is also required for short-term plasticity for language learning [10]. Still, how much infants actually utilize emotional feedback in their language learning is not precisely understood.

From computational simulation point of view, it seems that a primitive ability to perceive non-linguistic satisfaction from the caregiver can be utilized. However, the existing models utilizing caregiver feedback are mostly limited to the learning of speech production (see [1] and references therein). However, it would be also relevant to investigate how feedback may affect perceptual re-organization at the general auditory and linguistic levels.

3. Ability to perceive visual world

The meanings of words emerge from grounding of the auditory patterns to some external referents or concepts. For early LA, the simplest cases of grounding correspond to the linking of words into concurrently perceived and salient external objects and events. But what kind of perceptual processing can be assumed outside the auditory domain for a learner in the stage of learning its first words, corresponding to an infant of age approximately 6–12 months? Is it plausible to use discrete representations of the visual world to represent possible word referents, as in [11-13], or is it necessary to perform actual visual processing from real visual sensors such as in [14]?

What is known about infant object perception is that three to five-month-old infants are already able to perceive visual objects as separate complete entities having specific spatial boundaries [15], and much of this ability develops during the four first months of the infants' lives [2]. The early grouping of the visual scene into separate objects is mainly based on spatial segregation of connected surface properties and the common movement pattern of the elements of an object when the object or the viewpoint is moving. Infants are also able to perceive completeness of partly hidden objects and retain the identity of an object that moves temporarily out of view. They also perceive integrated representations of partially occluded objects at the age of five months ([15] and the references therein). Infants are known to form categories for variable visual percepts based on shared features inside category members, but the categorization principles and their change with age are not yet well understood [16]. As an example, infants under the age of six months are already able to categorize visually presented non-human animals into categories of different species [17,18].

It is likely that the conceptualization of the surrounding physical world truly starts only at the onset of active exploration and manual manipulation that begins once the infant has learned to sit and move around autonomously [2,19], normally corresponding to the age of 6–7 months. It is also likely that the development of visual and haptic perception is not strongly coupled to the development in auditory perception. This is supported by the numerous findings that innately deaf people do not show any kind of impairment in cognitive capability outside the language faculty, but perform at normal levels in non-verbal IQ tests, including tests that measure visuospatial skill [20]. Moreover, with sufficiently early exposure to spoken or sign language, the proficiency of deaf subjects in written language skills is comparable to the population with normal hearing [20].

As for the implications in terms of computational models of LA, one may assume that the LA agent is able to perceive the visual world as a collection of distinct objects and actors without essential loss in experimental plausibility, as is done in the majority of the existing computational studies (e.g., [1, 11-13]). Evidence is definitely not sufficient to rule out the possibility that low-level unbound visual features would not have any interaction with concurrent auditory processes, but as long as the visual processing itself is not the center of interest, it may be justified to assume that the agent is aware how visual features make up whole objects, and how same objects can occur in slightly different situations. It also seems evident that infants are able to cluster similar objects, such as different colored balls or different animals from a same species, into abstract categories, but this processing is strictly limited to perceivable features at first and not made according to more complex ontological constructions available to adults. However, caution should always be used in formulation of experiments with simulated referents, since it is not axiomatic that different types of referents such as actions, objects, animates and, e.g., adjectives acquire similar and distinct conceptual representations in the minds of the infants.

4. Ability for shared attention

Attention is considered to be an important factor in learning, since it allows the learner to focus on specific aspects of the environment at a time. In LA simulations, attention is typically considered as a mechanism that selects a subset of all possible visual objects (referents) in the immediate surroundings to be processed simultaneously with the language input. In many experiments (e.g., [11-13,21]; see also [1] and references therein), it is also assumed that the caregiver almost always speaks an utterance that concerns the objects and events that are jointly attended by both the caregiver and the learner. But how well do young infants actually follow attention?

One source of evidence that caregiver gaze direction modulates infant learning behavior comes from the event-related potential (ERP) study by Reid, Striano, Kaufan and Johnson [22]. They showed that once the infants were familiarized with sequences of two object images simultaneously with their caregiver always attending to one of the objects, the neural responses to the objects fixated by the caregiver were found to be different from the responses for non-attended objects. Behavioral experiments show that infants of age six months are able to follow an adult's gaze in the correct direction [23], but they often fixate on the first object seen on the path towards the target direction [24]. By 9-12 months of age, the infants seem to be able to attend reliably to the correct target fixated by the caregiver, but only if it is located within their visual fields [24-27]. At 12 months of age, the infants are also able to infer the focus of the attention of an adult when the adult is looking at an object behind a barrier blocking the infant's view of the object [28]. When caregiver attention is signified by a head turn gesture, the 12-month-old infants are also able to converge on the same focus of attention even outside the current visual field [29]. At 24 months of age, children assume that a novel name of an object is related to the object that the adult is concurrently

looking at instead of other objects that may be perceptually more salient, and even when the saliency is suddenly enhanced during the naming process by lighting up the object [30]. The attentional convergence also correlates with the ability to learn names for the objects at the age of six and eight months [9]. Finally, there is evidence that early attentional development is not strongly coupled to the normally concurrently developing ability to perceive spoken language, since congenitally deaf children show normal attentional engagement at the ages of six, nine and twelve months [31].

In general, the research suggests that the infants around the age of 9 to 12 months can follow the visual attention of the caregiver, but the convergence is modulated by the signifying gestures of the caregiver and the location of the target. Younger infants have trouble following attention outside their current field of vision, but this problem is alleviated as they approach the age of one year. It has been proposed that the emergence of and experience with motor skills enables more comprehensive understanding of the spatial relationships between objects and agents in the environment, also enabling following attention to currently unseen aspects of the environment [25]. This is also in line with what is known about the ability to perceive the environment (see the previous section). Interestingly, the ability for having a comprehensive joint attention with the caregiver develops at the approximately same age as the first tokens emerge in the receptive vocabulary of the learner. The comprehension and production of language has also been shown to correlate highly with the engagement in joint attention for 9 to 15-month old infants [32,33], suggesting that the attentional mechanisms play an important role in language learning.

For computational simulations concerned mainly with language learning, a plausible assumption seems to be that infants of approximately 12 months have the basic mechanism to follow the attention of the caregiver to a degree that allows them to separate objects and actions relevant to the present interaction from those environmental variables that are not in the focus of the caregiver. Assuming also the ability to form persisting conceptual representations of the objects and actors in the environment (section 3), the representation of the concurrent visual attended context as a set of categorical percepts is not entirely implausible simplification. Naturally, the way that the simulated visual percepts are represented internally in a computational system has to be specified separately for each study and cannot be addressed here in a general manner.

5. Cross-situational learning

Cross-situational learning (XSL; [34,35]) is not so much a constraint, but a possible solution to the word-referent problem expressed by Quine [36]: *how does the listener know what the novel word refers to*? The basic principle of XSL is that given multiple exposures to a word simultaneously with a number of possible visually perceived referent objects, the referential ambiguity becomes gradually solved since the co-occurrence probability of the word and the correct referent is higher than with the other candidate referents. As long as there are sufficiently many exposures, and as long as the word and its referent occur together at above chance level, the learning succeeds even if there is always an arbitrary sized subset of words and word referents in each communicative situation. In practice, the convergence of the attention between the infant and the caregiver helps to prune the group of possible referents to a relatively concise set for each interaction situation, especially when other basic principles of referential learning are taken into account, including the preference for naming of whole objects instead of their parts [37,38] and avoidance of multiple names for a single object (mutual exclusivity principle; [39,40] but see also [33] for criticism against innate word learning constraints). The XSL is also a generalization from the nameless category principle (N3C) of Golinkoff et al. [38] that states that a novel word is always first mapped to a novel object instead of any familiar object present in the scene. This is especially the case when cross-modal distributions need to be mutually consistent, e.g., when p(*referent*|*word*) and p(*word*|*referent*) yield the highest values for the same word-referent pair.

There is behavioral evidence that infants are sensitive to the cross-situational statistics in referential learning for novel words and objects [41], and also that adults are sensitive to fine-grained co-occurrence probabilities of multiple referential candidates [42,43]. Mathematical analysis of the learnability of words with the XSL using different amounts of referential uncertainty and with different sized vocabularies suggests that the mechanism may allow acquisition of large vocabularies in a reasonable time [44]. The study of the different XSL-based learning strategies in [43] suggests that adult listeners use a pure eliminative strategy (drop out unlikely referential candidates early) to figure out word meanings if the referential uncertainty is low, but use a more comprehensive frequency-based ranking for learning situations with high referential uncertainty ([43]; cf. [42]).

What the findings in XSL suggest is that a cognitive mechanism exists that is able to keep track of typical contexts in which specific auditory patterns occur and that the association strengths between the patterns and contextual variables seem to be driven by the frequencies of their co-occurrences. The idea of XSL-based lexical learning is greatly supported by the findings that infants are already able to perceive world as a spatiotemporal collection of discrete objects (section 2), and are able to share attention with their caregiver (section 3). In the existing computational studies of LA from continuous speech, the XSL principle has already been implicitly or explicitly used successfully in order to find the correct mapping between multiple acoustic patterns and multiple visual referents [1, 11-13]. Also, XSL can be utilized to learn synonymy and equivalence between acoustically different patterns that occur in similar functional contexts [45].

6. Conclusions

The complexity of computational simulations of LA inevitably becomes very high as they move towards more comprehensive descriptions of the process that necessarily integrate multiple developmental stages in order to understand the interplay of perception of speech, lexical acquisition, semantics, grammar, and speech production. In order to limit the complexity and in order to better understand the limitations of the studied models, attention must be paid to the assumptions that can be made regarding the other cognitive processes that are excluded from or taken for granted in the analysis. The observations made in the current review are not claimed to be conclusive so that the discussed processes are understood and should not deserve special modeling attention per se. Instead, the presented knowledge is meant to provide a rough skeleton for computational studies upon which more detailed simulations can be designed.

7. Acknowledgements

This research was funded by the Finnish Graduate School of Language Studies (Langnet).

8. References

[1] Räsänen, O., "Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions", *Speech Communication*, in press.

[2] Johnson, S. P., "Visual development in human infants: Binding features, surfaces, and objects", *Visual Cognition*, 8, 565–578, 2001.

[3] Montague, D. P., & Walker-Andrews, A. S., "Peekaboo: A New Look at Infant's Perception of Emotion Expressions", *Developmental Psychology*, 37, 826–838, 2001.

[4] Mastropieri, D., & Turkewitz, G., "Prenatal Experience and Neonatal Responsiveness to Vocal Expressions of Emotion", *Developmental Psychobiology*, 35, 204–14, 1999.

[5] Lasky, R. E., & Williams, A. L., "The Development of the Auditory System from Conception to Term", *NeoReviews*, 6, 141–152, 2005.

[6] Nazzi, T., Bertoncini, J., & Mehler, J., "Language discrimination by newborns: Toward an understanding of the role of rhythm", *J. Exp. Psych.: Human Perception and Performance*, 24, 756–766, 1998.

[7] Grossman, T., "The development of emotion perception in face and voice during infancy", *Restorative Neurology and Neuroscience*, 28, 219–236, 2010.

[8] Parncutt, R., "Prenatal and infant conditioning, the mother schema, and the origins of music and religion", *Musicae Scientiae*, 13, 119–150, 2009.

[9] Gogate, L. J., Bolzani, L. H., & Betancourt, E. A. (2006). Attention to Maternal Multimodal Naming by 6- to 8-Month-Old Infants and Learning of Word-Object Relationships. *Infancy*, 9, 259–288.

[10] Kuhl, P. K., Tsao, F-M., & Liu, H-M., "Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning," *Proc. National Academy of Sciences*, 100(15), 9096–9101, 2003.

[11] ten Bosch, L., Van hamme, H., & Boves, L., "Discovery of words: Towards a computational model of language acquisition", In F. Mihelič and J. Žibert (Eds.), *Speech Recognition: Technologies and Applications* (pp. 205–224). Vienna: I-Tech Education and Publishing KG, 2008.

[12] Räsänen, O., & Laine, U. K., "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences", *Pattern Recognition*, 45, 606–616, 2012.

[13] Aimetti, G., "Modelling early language acquisition skills: Towards a general statistical learning mechanism", *Proc. of EACL-2009-SRWS*, Athens, Greece, 1–9, 2009.

[14] Roy, D., "Grounded Spoken Language Acquisition: Experiments in Word Learning", *IEEE Trans. Multimedia*, 5, 197–209, 2003.

[15] Spelke, E. S., "Principles of Object Perception", *Cognitive Science*, 14, 29–56, 1990.

[16] Marechal, D., & Quinn, P. C., "Categorization in Infancy", *Trends in Cognitive Sciences*, 5, 443–450, 2001.

[17] Eimas, P. D., & Quinn, P. C., "Studies on the formation of perceptually based basic-level categories in young infants", *Child Development,* 65, 903–917, 1994.

[18] Oakes, L. M., & Ribar, R. J., "A comparison of infants' categorization in paired and successive familiarization", *Infancy,* 7, 85–98, 2005.

[19] Piaget, J. "*The construction of reality in the child*", New York: Basic Books, 1954.

[20] Mayberry, R. I., "Cognitive development in deaf children: the interface of language and perception in neuropsychology", In S. J. Segalowitz. & I. Rapin (Eds.), *Handbook of Neuropsychology, 2nd Edition, Vol. 8, Part II*. Elsevier, Amsterdam, 2002.

[21] ten Bosch, L., Räsänen, O., Driesen, J., Aimetti, G., Altosaar, T., & Boves, L., "Do Multiple Caregivers Speed up Language Acquisition?", *Proc. of Interspeech'09,* Brighton, England, 704–707, 2009.

[22] Reid, V. M., Striano, T., Kaufman, J., & Johnson, M. H., "Eye gaze cueing facilitates neural processing of objects in 4-month-old infants", *NeuroReport*, 15, 2553–2555, 2004.

[23] D'Entremont, B., Hains, S. M. J., & Muir, D. W., "A demonstration of gaze following in 3- to 6-month-olds", *Infant Behavior and Development*, 20, 569–572, 1997.

[24] Butterworth, G. E., & Jarret, N., "What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy", *British Journal of Developmental Psychology*, 9, 55–72, 1991.

[25] Flom, R., Deák, G. O., Phill, C. G., & Pick, A. D., "Nine-month-olds' shared visual attention as a function of gesture and object location", *Infant Behavior & Development*, 27, 181–194, 2004.

[26] Brooks, R., & Meltzoff, A. N., "The importance of eyes: how infants interpret adult looking behavior", *Developmental Psychology*, 38(6), 958–966, 2002.

[27] Woodward, A. L., "Infants' developing understanding of the link between looker and object", *Developmental Science*, 6, 297–311, 2003.

[28] Moll, H., & Tomasello, M., "12- and 18-month-old infants follow gaze to spaces behind barriers", *Develop. Sci.*, 7, F1–F9, 2004.

[29] Deák, G., Flom, R., & Pick, A. D., "Effects of gesture and target on 12- and 18-month-olds' joint visual attention to objects in front of or behind them", *Developmental Psychology*, 36, 511–523, 2000.

[30] Moore, C., Angelopoulos, M., & Bennet, P., "Word learning in the context of referential and salience cues", *Developmental Psychology*, 35, 60–68, 1999.

[31] Spencer, P. & Waxman, S., "Joint attention and maternal attention strategies: 9, 12 & 18 months", In *Maternal responsiveness and child competency in deaf and hearing children*. Final Report, Grant H023C1077, OSERS, US, Department of Education, 1995.

[32] Carpenter, M., Nagell, K., & Tomasello, M., "Social cognition, joint attention, and communicative competence from 9 to 15 months of age", *Monographs of the Society for Research in Child Development*, 63(4), 1–143, 1998.

[33] Tomasello, M., "The Social-Pragmatic Theory of Word Learning", *Pragmatics*, 10, 401–413, 2000.

[34] Pinker, S., "*Learnability and cognition: The acquisition of argument structure*", Cambridge, MA: MIT Press, 1989.

[35] Gleitman, L. R., "The structural sources of verb meanings. *Language Acquisition*", 1, 3–55, 1990.

[36] Quine, W. V. O., "*Word and object*", Cambridge, MIT Press, 1960.

[37] Macnamara, J., "The cognitive basis of language learning in infants", *Psychological Review*, 79, 1–13, 1972.

[38] Golinkoff, R. M., Mervis, C. B., & Hirsh-Pasek, K., "Early object labels: the case for a developmental principles framework", *Journal of Child Language*, 21,125–155, 1994.

[39] Markman, E. M., & Wachtel, G. F., "Children's use of mutual exclusivity to constrain the meaning of words", *Cognitive Psychology*, 20, 121–157, 1988.

[40] Markman, E. M., "Constraints on word meaning in early language acquisition", *Lingua*, 92, 1–4, 1994.

[41] Smith, L. B., & Yu, C., "Infants rapidly learn word-referent mappings via cross-situational statistics", *Cognition*, 106, 1558–1568, 2008.

[42] Vouloumanos, A., "Fine-grained sensitivity to statistical information in adult word learning", *Cognition*, 107, 729–742, 2008.

[43] Smith, K., Smith, A. D., & Blythe, R. A., "Cross-Situational Learning: An Experimental Study of Word-Learning Mechanisms", *Cognitive Science*, 35, 480–498, 2011.

[44] Smith, K., Smith, A. D., Blythe, R. A., & Vogt, P., "Cross-situational learning: a mathematical approach", *Proc. Third International Workshop on the Emergence and Evolution of Linguistic Communication*, Rome, Italy, 31–44, 2006.

[45] Räsänen, O., "Context induced merging of synonymous word models in computational modeling of early language acquisition", *Proc. ICASSP'2012*, Kyoto, Japan, pp. 5037-5040, 2012.