Language Acquisition between Computerized Agents

Benjamin Han

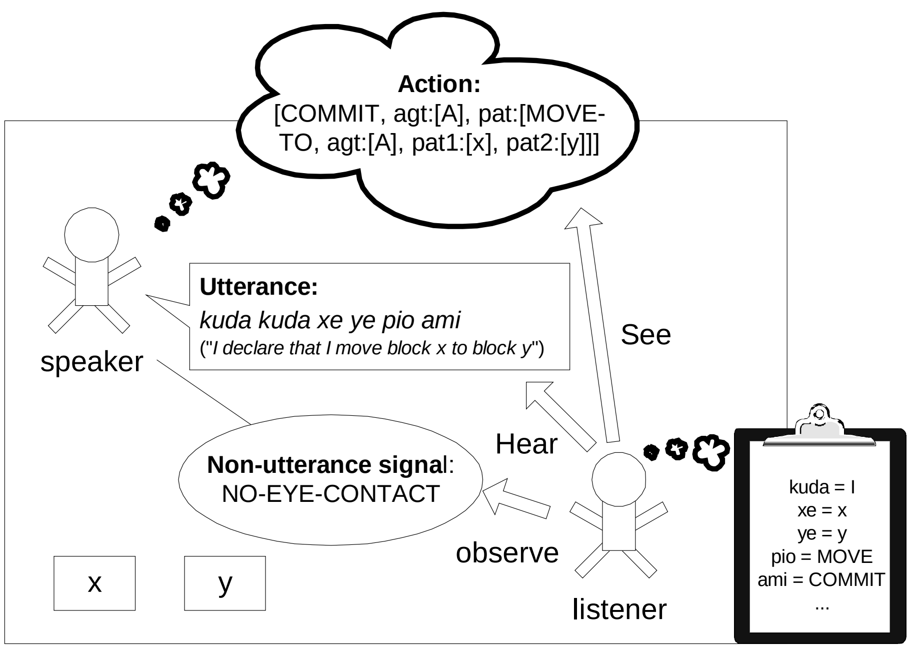
Department of Computer Science and Engineering,
State University of New York at Buffalo, NY 14228

dh6@acsu.buffalo.edu

Motivation

Through out the most part of the human history, languages have been the only tools which can be used to describe and exchange experiences, knowledge and abstract concepts effectively and efficiently. With the complexity and subtlety of languages, it is an amazing feat that young children are able to pick up their mother tongue fairly quickly[1]. Although the exact cognitive process involved for language acquisition (LA) has yet been determined, there are facts emerging from numerous empirical studies.

In this paper, a computational model of LA in a controlled environment is proposed based on the work [14] and the findings in the empirical studies. To be more precise, I focus on the LA problem between two computerized agents 'living' in the same controlled environment. Both agents can hear the utterance and observe the behaviors of the other agent, and their sensory systems are capable of translating the same event into exactly the same internal representation. Thus the LA problem is reduced to finding the correct mappings between the language of the speaker agent and the internal representation of the listener agent.[2] This is illustrated in the following figure.



---

[1] Most children will say their first words about one year after their birth.

[2] It is not required for the listener agent to have its own language – the basic idea is to find the mapping between the observed utterance and the *ground truth*, namely the internal representation used by the two agents. Both agents use the same internal representation, which serves as a "universal language" to 'link' between different languages.

The proposal serves two purposes. Firstly it provides a starting point on which we can ponder on exactly what elements, in addition to the obvious one, the utterance, are required in the process of LA. Secondly, for practical purposes, it is desired to have autonomous programs (computerized agents) to communicate with each other without explicitly hardwiring a communication protocol for every possible language. Given appropriate innate mechanisms an agent should have no difficulty picking up the utterance of the other one, through its observations and learning processes.

The rest of the paper is organized as follows. In Section 2 the characteristics of LA are explored by summarizing the relevant theoretical and empirical studies, thus serve as the fundamentals for the proposed computational model. In Section 3 the problem is formalized and the assumptions are stated. The learning process for the computerized agent is then proposed in Section 4 by giving an example. Finally the limitations of the model, together with the possible extensions, are discussed in Section 5.

Language Acquisition

Before 60s the ideas that language is learnt by mere observation and imitation without any hardwired language faculty was popular among empiricists/behaviorists. In particular Quine [11] suggested that children learn the meaning of a word by merely associating it with the salient properties of their perceptual experience when the word is used, without resorting to any ontological commitments. For example, when a newborn hears the word 'stick' at the time a stick is presented, she would interpret the word 'stick' as "a portion of stick experience", without taking the solidity of a stick into account. Similarly, the word 'water' together with the presence of water would be interpreted as "a portion of water experience." However, the theory had difficulty explaining how the difference between the quantifiers 'another' and 'more' as in "another stick" and "more water" could be appreciated, if the mind of infants, as Aristotle proposed, is a blank slate. To make learning possible, newborns must be able to at least discriminate linguistically meaningful signals categorically [9]. By using a creative approach for observing the attentiveness of the newborns[3], P. Eimas et al [5] showed that four-month-old babies are capable of distinguishing the syllable [ba] from [pa]. Later experiments conducted by Juscyzk and Derrah [8] further showed that the newborns perceive syllables as atomic elements. Beyond syllables, newborns demonstrated their abilities to distinguish other linguistically meaningful signals as well, such as prosody [10]. In addition to the innate ability for categorical perception, experiments showed that young children indeed have ontological commitments. In the experiments reported in [13] the young children are able to correctly identify similar objects by using different criteria depending on whether the object is solid object or nonsolid substance.

To account for the rapid pace a newborn masters language, Chomsky [4] suggested that part of the language faculty is innate, in particular, there exists a hardwired rule system (*universal grammar*) which serves as the base system for all human languages. Only a portion of the faculty is learnable, and it is through the interactions between the culture the newborn is in that the "parameter values" are fixed to form a particular human language. This proposed semi-innate system explains why young children could learn new words at an amazing rate [3], as well as why children in multilingual environment could acquire all of the languages at the same pace as those who are in monolingual environment[4].

Although auditory input plays an important role for a newborn to learn language, it is by no means the only medium exploited for the purpose of LA, for deaf children have proved that they are as good language learners as the normal ones. Therefore to model the LA process we should take advantage of the essential non-utterance media, such as gestures, facial expressions, behavior, etc, as well. The idea that language is learnt through the relevant context dates back to Wittgenstein [16], who defined the concept of "language game" as a combination of words and actions, attitudes and behavior, namely, the whole process of using

[3] Essentially the amplitude of the sucking behavior of a newborn is measured when a stimulus is presented to her (in this case it is sound). If the baby is capable of relating her sucking with the occurrence of the stimulus, the desire of being stimulated should make the newborn suck with higher amplitude. This behavioral change then is interpreted as a proof of the infant's ability to differentiate the different stimuli.

[4] The reasoning goes like this: assuming all newborns devote the same amount of time in sleeping, the amount of the perceptual experiences they receive should be the same. Therefore the babies in multilingual homes should have less exposure to each language, which implies, according to empiricists/behaviorists claims, that they should have learnt each language less well.

the language. The full understanding of language would have been impossible had the newborns been allowed to perceive only the acoustics of language.

Before trying to sum up the crucial dimensions which must be taken into account in order to realize the LA process in an artificial setup, let us consider the feasibility of LA. Empirical studies have shown that the ability to acquire language is to some extent independent from the general cognitive faculties. In [2] the language learning abilities of the children suffering from Williams syndrome, thus having very low IQs, are almost unaffected, while in [15] children with normal or superior IQs have difficulties acquiring language effectively. Although IQ as *the* measure of intelligence is still in debate, this indicates that in our mind the language faculty is isolated from at least some other modules, which are responsible for higher/lower IQ. The modularity of mind thus motivates the possibility of modeling LA even when the likelihood of realizing general AI is still in hot debate.

In summary, to successfully realize the LA process we need to not only replicate the essential innate system, but also preserve the crucial communication clues for the learning process. For the particular setup in this paper, for the former part we assume a common ontology between the two agents. Namely, both agents have the same internal representation of their surrounding environment, and they have the same set of *execution schemas*[5]. For the latter part, we assume that for each agent its intensionality of the utterance and a limited set of its emotion manifest themselves via digitized non-utterance signals, such as *eye-contact*. These will serve as important clues for learning the mappings between the word forms of various perfomatives and their respective internal representation.

Problem Formulation

In this section and the following one the particular problem addressed in this paper will be formulated and the solution proposed, based on the work in [14], and various empirical studies and hypotheses made in the previous section. Let us consider two computerized agents *A* and *B* living in the same artificial environment. Assume *A* has its own language and *B* has none, the problem is designing a learning procedure for *B* by which it is able to learn the meaning of *A*'s utterances – this includes both the lexical meanings and the grammar of *A*'s language. The environment is fairly simple – there are blocks with different colors/shapes scattered around, and agent *A* will do various actions on the blocks and at the same talks to agent *B*.

To be more precise, let $I = \{ (u_i, s_i, a_i) \mid u_i$ is the *i*-th utterance of *A*, s_i is the *i*-th non-utterance signals of *A*, and a_i is *A*'s *i*-th action that accompanies $u_i\}$ be the communication sequence from *A* to *B*, where

- Utterance: a list of words in *A*'s language. There are two restrictions imposed on the utterance. Firstly every utterance should contain one performative component, which can be either expressed in one word, or in a structure. For example, in utterance "*I declare I move block x on top of y*" the part "*I declare*" is a *commissive*. In this paper we only consider *commissive* and *directive* ("*I request you to move x on top of y*"). But the model could be extended given the formalizations of the semantics of the other performatives as done in [12]. The second restriction is that the underlying grammar of *A*'s language should be context-free. This arbitrary restriction allows agent *B* to infer the correct word ordering by assuming *A*'s language is governed by context-free grammar (CFG), and complies with Chomsky's notion of language faculty (see Section 2). Note it is possible to extend the model by allowing less restrictive grammar.

- Non-utterance signals: these are represented by a list of primitives, such as EC (*eye-contact*), NEC (*no-eye-contact*), SAT (*satisfied*), and USAT (*unsatisfied*). Whenever an agent speaks, it either establishes eye contact with the listener, or it does not. With this clue in the former case the listener then can infer the performative the speaker is using must be directed to the listener (*directive*), while in the latter case the utterance is either a broadcasting one or is directed to the speaker itself (*commissive*). If the listener has the ability to identify a directive utterance, it can then have an idea as to whether it successfully

[5] The same term *execution schema* (*x-schema* for short) is used in [1]. The basic idea behind the two usages are the same, namely, there are *innate concepts* about the common actions, such as *push* and *pull*.

fulfill the speaker's request by observing the SAT/USAT signal from the speaker. Essentially the primitives for non-utterance signals serve as non-linguistic communications necessary for LA.

- Action: this is represented in a tree-like structure, with the root node denoting one of the predefined execution schema, and the leaf nodes denoting the agent and the patient(s) of the action. For example, [MOVE-TOP, agt:[A], pat1:[x], pat2:[y]] denotes that agent *A* moves block *x* on top of block *y*. The possible execution schema include physical actions: MOVE, MOVE-TO, MOVE-ON, MOVE-BESIDE, WALK-TO, WALK-ON and WALK-DOWN, as well as performatives: COMMIT and DIRECT. The existence of these schemas constitutes the innate part of the language faculty of our computerized agents.

The goal here is to design a learning procedure for agent *B* so that it can learn a mapping *M* of (word, meaning) pairs, such that the size of *M* will be minimal. The minimality of *M* is simply a working hypothesis, and in time it must be justified empirically. Another important assumption is *compositionality* of semantics. The assumption firstly states that the meaning of a sentence (utterance) is composed from the meanings of the sentence constituents (words). Secondly from the reverse direction, a part of the sentence meaning should be contributed by the meaning of a single sentence constituent. Formally speaking, let *R* be the set of all actions observed so far, *P* be the set of all the nodes in all actions in *R*, and *C* be a constructor relation mapping from 2^P (power set of *P*) to *R*, i.e., $C \subseteq 2^P \times R$. For an utterance $u_i = (w_{i1}, ..., w_{im})$ then if $(w_{ij}, p_j) \in M$ for $j=1...m$, then $(\{p_1, ..., p_m\}, r_i) \in C$.

Intuitively, the learning proceeds as follows. Every time before *A* performs an action, it will utter a sentence, which is either a commissive or a directive statement, and exhibit non-utterance signals. Agent *B* will hear the utterance, perceive the non-utterance signals together with the actual action taken. It then tries to find the correct mappings *M* between the words in the utterance and the sub-structures in the action observed, taking the non-utterance signals into account. If *B* determines that *A*'s utterance is a directive statement, it then must try to satisfy whatever it believes that *A* asks it to do. In subsequent time steps *B* will learn whether he did is right or wrong from *A*'s non-utterance signals SAT or USAT, and adjusts its learnt mappings accordingly[6].

The particular formulation here, compared to the work done in [14], has what I believe an important twist for LA process. The various speech acts are taken into account in the actual learning process, and the interactivity in turns enables agent *B* to not only learn the lexical meanings passively, but also learn the correct grammar by experimenting (attempting to fulfill agent *A*'s request). The mappings between performative word forms and their meanings is achieved by the addition of non-utterance signals, which I believe completes a computer realization of Wittgenstein's conception about language games. In the following section it should be clear that non-utterance signals give important hints about the particular speech act involved in an utterance.

Learning Process

Let us illustrate the learning process by an example. The language of agent *A* is specified in the following table. The semantics should be clear once the example is presented. Note I deliberately design this seemingly gibberish-like language to defy the language intuitions English speakers have. Also note some of the words are polysemous – they have more than one meaning. The learning procedure described will allow agent *B* to learn polysemous words. In fact, it can be illustrated that synonyms can also be learnt by this procedure.

Now imagine *A* issued the first communication to *B*, which is shown below in (*utterance, non-utterance signals, actions*) triple:

("*kuda kuda xe ye pio ami*", (NEC), [*?p*, agt:[A], pat:[MOVE-TO, agt:[A], pat1:[x], pat2:[y]]])
(English translation: "*I declare that I move block x to block y*")

[6] The obligations that A must make an utterance exactly corresponding to its action, B must try to understand what A is talking about, and B must try to fulfill A's request are implied, or hardwired. In many ways the learning protocol reminds one of Grice's *conversational maxims* [6].

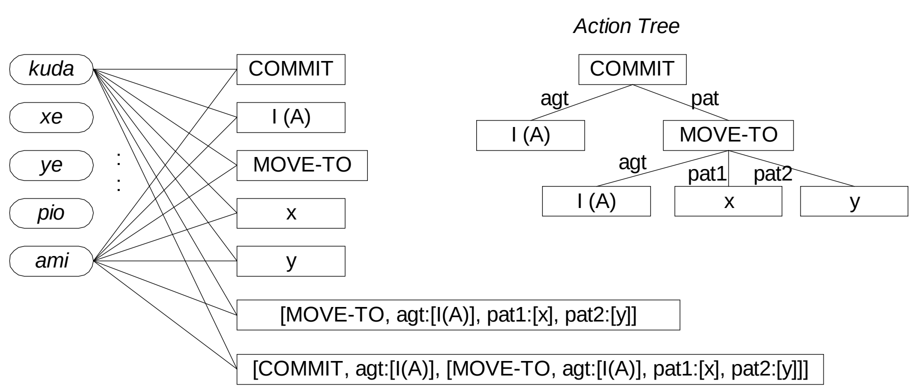
Note in *B*'s perception it is still unknown which performative *A* is using in this particular communication, hence we put a variable *?p* in the action element. Now according to the innate mappings between non-utterance signals and performatives, *B* knows that *?p* must be COMMIT because *A* did not make eye contact with it (NEC) during the communication (*A* was talking to either itself or everyone, in the latter case the utterance is not meant to be directed to only *B*), and the only remaining possibility, DIRECT, does require an EC. Hence the action is revised to

[COMMIT, agt:[A], pat:[MOVE-TO, agt:[A], pat1:[x], pat2:[y]]].

Vocabularies			Syntax
	Word	Meaning	S → *kuda* S′ P
Indexical (I)	*tuda*	You	S′ → I Obj V
	kuda	I (the speaker)	Obj → N | N Obj
Noun (N)	*xe*	x	
	ye	y	
	ze	z	
Verb (V)	*pio*	MOVE MOVE-TO MOVE-ON MOVE-BESIDE	
	pek	WALK WALK-TO WALK-ON WALK-DOWN	
Performatives (P)	*ami*	COMMIT	
	tuko	DIRECT	

Since the semantics of performative COMMIT, as formulated in [12], requires that the object of commitment to be satisfied[7], agent *B* determines that *A*'s utterance must conform to what it observes in *A*'s action. The remaining problem, then, is finding the correct mappings between the words in *A*'s utterance and the sub-structures in *A*'s action.

Without more clues, at this stage *B* can only have a most general guess of the meanings of each word, namely, for each word the possible meaning is any sub-tree of the observed action tree. This is illustrated below as a bipartite graph, in which a link between a word and a sub-tree of *A*'s action denotes one possible mapping. Note although in *B*'s eyes the agents of all actions are *A*, we assume here that *B* knows that when *A* makes a commitment to itself, it will address *A* (itself) using indexical "*I*" instead of third-person address *A*. This enables *B* to correctly learn the meaning of *kuda* at the later stages.

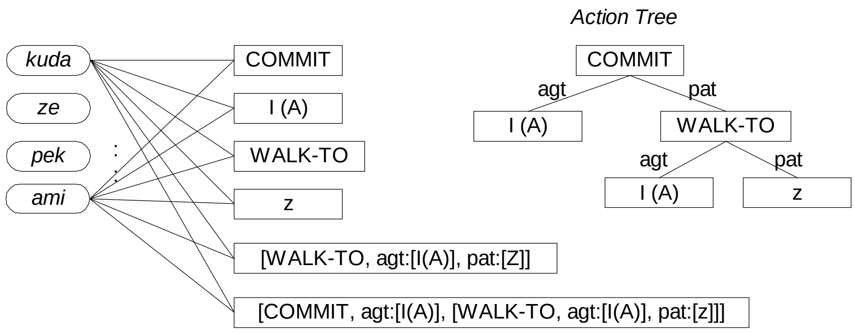


Now it is time for *B* to observe the second communication from *A*:

("*kuda kuda ze pek ami*", (NEC), [COMMIT, agt:[A], pat:[WALK-TO, agt:[A], pat:[z]]])
(English translation: "*I declare that I walk to block z*")

[7] In [7] the author described the concept of *wholehearted satisfaction* as the non-contingency satisfaction of a request.

Note we already put COMMIT in place by following similar reasoning illustrated in learning the first communication. Assuming this is the only communication *B* receives from *A*, we have the following bipartite graph denoting the possible mappings between the words and the meanings:



Now let us compare the two bipartite graphs *B* has so far. Consider word *kuda*. Since we have the mapping (*kuda*, I(A)) in both graphs, the mapping has a *coverage* 100%. Similarly (*kuda*, COMMIT) has 100% coverage, but (*kuda*, WALK-TO), (*kuda*, z), (*kuda*, [WALK-TO, agt:[A], pat:[z]]) and (*kuda*, [COMMIT, agt:[A], pat:[WALK-TO, agt:[A], pat:[z]]]) all have 50% coverage. An important heuristic at this stage is *choosing the mapping which has the largest coverage*. For *kuda* we have two possibilities, let us consider the first one: (*kuda*, I(A)). If this is the case, then it must be the case that the mapping (*ami*, COMMIT) is true, since by the heuristic we just describe the only two possibilities for word *ami* is (*ami*, COMMIT) and (*ami*, I(A)). By similar reasoning another possible pair of mappings is (*kuda*, COMMIT) and (*ami*, I(A)). Essentially, the choices of one mapping for a word will constrain the choices of the mapping for another word appearing in the same sentence, because of the assumption of semantics *compositionality*.

Now consider the third communication from *A* to *B*:

("*kuda tuda ze pek tuko*", (EC), [DIRECT, agt:[A], pat:[*?a*, agt:[B], pat:[*?p*]])
(English translation: "*I request you to walk to block z*")

Note now *A* has eye contact (EC) with *B*, which signals the use of performative DIRECT, and by obvious reason the agent of DIRECT must be *A*. This enables *B* to conclude that the mapping (*kuda*, I(A)) must be true, since the other possible mapping (*kuda*, COMMIT) no longer exists in this communication. This in turn implies that the mapping (*ami*, COMMIT) must be true.

However, agent *B* now faces a novel problem – it does not know what *A* asks it to do. The only thing it knows is that the agent of the required action must be itself (*B*). Furthermore, even if *B* knows the correct mappings for all the words in *A*'s utterance, it still has to learn the correct grammar to map the word orderings into the correct action tree. To simplify the problem, let us assume at this stage *B* has learnt the correct meanings of all the words, i.e., it knows the following additional mappings: (*tuda*,You(B)), (*ze*, z), (*pek*, WALK-TO) and (*tuko*, DIRECT) by observing many commissive communications from *A* to *B*[8]. Since all execution schemas are innate, *B* must already know, for example, that the action WALK-TO requires two arguments, i.e., one *agent* and one *patient*. Thus the possible grammar for WALK-TO consists 3!=6 permutations:

1: *S'* $\rightarrow$ agent patient *pek*, 2: *S'* $\rightarrow$ agent *pek* patient, 3: *S'* $\rightarrow$ *pek* agent patient,
4: *S'* $\rightarrow$ *pek* patient agent, 5: *S'* $\rightarrow$ patient agent *pek*, 6: *S'* $\rightarrow$ agent patient *pek*,

of which only 1 and 5 are appropriate here, since we already know the correct mappings of the words. To break the tie, recall from the second communication that we have a sub-utterance "*kuda ze pek*". If we

[8] If not, *B* can simply choose to do nothing at this stage but wait for the next commissive communication in order to learn more mappings before it decides to try to fulfill *A*'s request.

choose 1 as the true grammar, the pattern has 100% coverage. But if we choose 5, the coverage will be smaller. Hence by the similar heuristic used to determine the best word-meaning mapping, we determine the best grammar to account for the observations is 1.

In practice when *B* receives a directive communication it may not know the correct word meaning, or may not have conclusive statistics to support the correct grammar. In such cases *B* simply randomly choose one candidate grammar and generate the action tree accordingly. In some subsequent time step it will observe the *A* communicates with it by either SAT or USAT non-utterance signal – this tells *B* if the candidate is the truth or not.

The rest of the LA process proceeds in the similar fashion described in the previous three communications. During the process *B* could even initiate the communication and observe *A*'s actions and adjust its learnt word meanings and grammar accordingly.

Discussions

In this paper the LA process of computerized agents living in a controlled environment has been modeled as a computational process. From the computational point of view, various assumptions have been made to make the problem tractable. These include: (1) the agents share the same ontology (2) the agents have innate mappings between non-utterance signals and the semantics of performatives, (3) the grammar of the target language is assumed to be CFG, and (4) the meanings of utterance are composed by their constituents and vice versa. Although the problem is discussed in an artificial setup, once we have empirical proofs supporting each of the four assumptions, it is not impossible to replicate the LA processes between two humans.

There are several possible extensions to this proposal. Firstly, the mappings between words and their meanings, together with the grammar, has been computed by a deterministic representation. Namely, in an utterance a word either has a particular meaning or does not have it. It is possible to extend the model so that the word-meaning mapping as well as the grammar are determined by probabilistic models. This might allow the ambiguity of language to be represented in the proposed model. Another possible extension is to generalize the learning process to a generic model-based reasoning problem, so that each prediction of the word meaning and grammar rules may or may not contradict to the new observations, and the existing models can then be refined accordingly.

Finally, note that the limited language discussed in this paper does not cover the notions about time (*past*, *future*, etc), modal truth, belief and the other more subtle concepts natural language could carry. To accommodate these concepts the model needs to be extended by using a more powerful action representation, such as temporal and modal logic.

Reference

[1] D. Bailey, J. Feldman, S. Narayanan, and G. Lakoff. *Modeling Embodied Lexical Development*, ICSI, UC Berkeley, 1997.

[2] U. Bellugi, S. Marks, A. Bihrle, and H. Sabo. Dissociation between Language and Social Functions in Williams Syndrome. In K. Mogford and D. Bishop, eds., *Language Development in Exceptional Circumstances*. Churchill-Livingstone Inc., 177 – 189, 1988.

[3] S. Carey. The Child as Word Learner. In M. Halle, J. Bresnan, and G. A. Miller, eds., *Linguistic Theory and Psychological Reality*, MIT Press, 1978.

[4] N. Chomsky. On the Nature, Use, and Acquisition of Language. In A. I. Goldman, editor, *Readings in Philosophy and Cognitive Science*, MIT Press, 1993.

[5] P. D. Eimas, E. R. Siqueland, P. Jusczyk, and J. Vigorito. Speech Perception in Infants. In *Science* 171: 303 – 306, 1971.

[6] H. P. Grice. Logic and Conversation. In P. Cole and J. Morgan, eds., *Syntax and Semantics*, vol. 3: *Speech Acts*. New York: Academic Press, 41 – 58, 1975.

[7] C. L. Hamblin. *Imperatives*, 151 – 157, Blackwell, 1987.

[8] P. W. Jusczyk and C. Derrah. Representation of Speech Sounds by Young Infants. In *Developmental Psychology* 23: 648 – 654, 1987.
[9] A. M. Liberman, K. S. Harris, J. A. Kinney, and H. Lane. The Discrimination of Relative-Onset Time of the Components of Certain Speech and Nonspeech Patterns. In *Journal of Experimental Psychology* 61: 379 – 388, 1961.
[10] J. Mehler, J. Bertoncini, M. Barrière, and D. Jassik-Gershenfeld. Infant Recognition of Mother's Voice. In *Perception* 7: 491 – 497, 1978.
[11] W. V. Quine. *Word and Object*. MIT Press, 1960.
[12] M. P. Singh. A Semantics for Speech Acts. In M. N. Huhns, M. P. Singh, and L. Gasser, eds., *Readings in Agents*, 458 – 470, Morgan Kaufmann Publish., 1998.
[13] N. N. Soja, S. Carey, and E. S. Spelke. Ontological Categories Guide Young Children's Inductions of Word Meaning. In A. I. Goldman, editor, *Readings in Philosophy and Cognitive Science*, MIT Press, 1993.
[14] C, A, Thompson and R. J. Mooney. *Lexical Acquisition: A Novel Machine Learning Problem*. Dept. of Computer Sciences, UT Austin, 1996.
[15] H. K. J. van der Lely. Language and Cognitive Development in a Grammatical SLI boy: Modularity and Innateness. In *Journal of Neurolinguistics* 10: 75 – 107, 1997.
[16] L. Wittgenstein. *Philosophical Investigations*. Blackwell, 3rd ed., 1968.