Preparing to Talk: Interaction between a Linguistically Enabled Agent and a Human Teacher

Caroline Lyon, Chrystopher L. Nehaniv and **Joe Saunders**
Adaptive Systems Research Group
University of Hertfordshire
Hatfield, AL10 9AB, United Kingdom
Email: {C.M.Lyon, C.L.Nehaniv, J.1.Saunders}@herts.ac.uk

Abstract

As a precursor to learning to use language an infant has to acquire preliminary linguistic skills, including the ability to recognize and produce word forms without meaning. This develops out of babbling, through vocal interaction with carers. We report on evidence from developmental psychology and from neuroscientific research that supports a dual process approach to language learning. We describe a simulation of the transition from babbling to the recognition of first word forms in a simulated robot interacting with a human teacher. This precedes interactions with the real iCub robot.

1 Introduction

When an infant utters his first meaningful words he has already passed through earlier stages preparing him for the acquisition of linguistic skills. The perception and production of words and phrases with referential meaning first require the ability to recognize word forms, to segment an acoustic stream of speech, (Vihman, dePaolis, and Keren-Portnoy 2009; Clark 2009; de Boisson-Bardies 1999). An infant passes through developmental stages of vocalization from canonical babbling to babbling that is biased towards the productions of his carer, moving on to the production of words that he hears, words that initially have no referential meaning.

The work described in this paper models the interaction between a human teacher and a simulated robot analogous to a child about 6 to 14 months old. We investigate how the language knowledge of this synthetic agent might develop to the point where it can recognize and produce word forms. This work precedes current investigations into real time interactions with the real iCub robot, demonstrating the underlying concepts. It is part of a much larger research programme aiming to facilitate the acquisition of linguistic skills by robots, undertaken by the ITALK project (see Acknowledgements).

We take a constructionist approach to language learning, following Tomasello (2003) and Bloom (2002), also inspired by Roy and Pentland (2002). In the wider picture ongoing work aims to enable the robot to ground speech patterns and gestural actions of a human with action, visual, proprioceptive, and auditory perceptions. It includes experiments on learning to name shapes and objects, where a human interacts with the humanoid robot Kaspar which has the appearance of a small child[1].

Learning lexical semantics would in reality overlap with word form acquisition, but we are investigating these processes separately initially, in order to understand each strand better. An account of ongoing work on acquiring lexical semantics is given in Saunders et al. (2009). In order to develop this stage it is necessary to recognize some word forms so that salient elements in the speech stream can be detected.

Experimental Context

Our work is based on the hypothesis that a robot or synthetic agent can acquire preliminary linguistic skills through interaction with a human teacher. The human side of the conversation in this simulation is taken from a corpus of dialogs between naive human teachers and Kaspar. This human speech from embodied human-robot interactions is used to evoke responsive utterances and drive word form learning by a simulated robot called "Lesa" (Linguistically Enabled Synthetic Agent), a synthetic counterpart of Kaspar.

In our research we are inspired by language acquisition in human infants, noting the great variation between individuals, but not bound to social and biological realism. The scenario for our experiments is a situation in which Kaspar and Lesa are taught about various shapes. The teachers are participants, not working on the project, who are asked to converse with Kaspar using their own spontaneous language, see Section 4.

We start from the basis that there is continuity between babbling patterns and early word productions (Vihman 1996). We describe experiments to simulate some aspects of these early processes, starting from the "landmark event" when canonical syllables emerge (ibid, p.118).

Initially Lesa's responses are analogous to canonical babbling, a string of random syllables. As the dialog progresses Lesa's output becomes biased towards what it has heard from its teacher. The probability of its producing a given syllabic form increases as it hears the syllable in its teacher's

[1] Kaspar is a minimal expressive child-sized humanoid robot developed by the University of Hertfordshire Adaptive Systems Research Group specifically for human-robot interaction. See Dautenhahn et al. (2009) for design details and rationale. Kaspar2 is used in the work described here

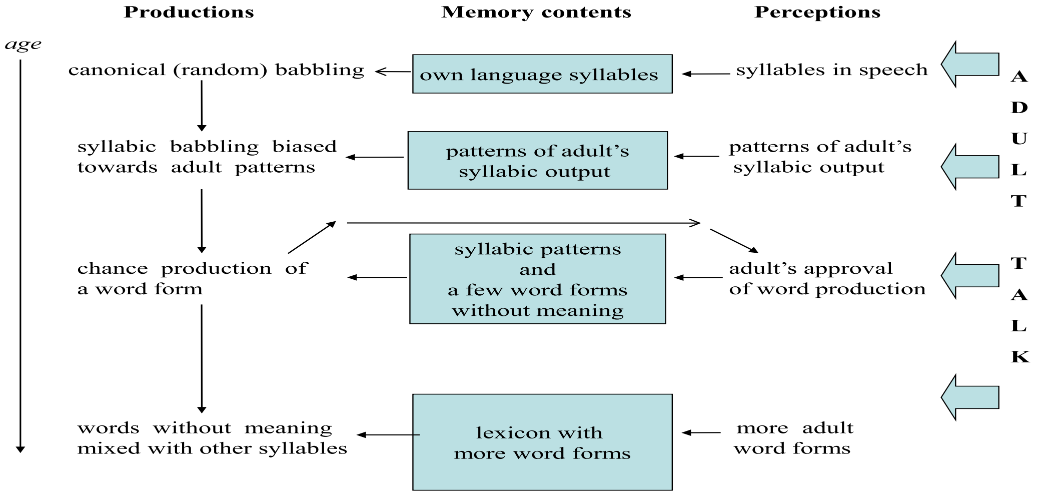
Productions
Memory contents
Perceptions
age
canonical (random) babbling
own language syllables
syllables in speech
syllabic babbling biased towards adult patterns
patterns of adult's syllabic output
patterns of adult's syllabic output
chance production of a word form
syllabic patterns and a few word forms without meaning
adult's approval of word production
words without meaning mixed with other syllables
lexicon with more word forms
more adult word forms
A D U L T T A L K

Figure 1:
Overview of the process of learning word forms through interaction with a teacher

speech. Eventually it will by chance produce a monosyllabic word that the human is trying to teach: "square", "sun", "moon" etc. when the teacher will signal approval. Polysyllabic words come later.

2 Background

Over the last two decades substantial evidence has been accumulated to show how very young, pre-linguistic infants have necessary precursors for acquiring linguistic skills. See for instance, among numerous examples, (Morgan and Demuth 1996) on the perception of prosodic information and (de Boisson-Bardies 1999; Clark 2009) on categorical perception.

Categorical perception is the observed phenomenon that humans take an analog speech signal and map it onto discrete phonemes. Before about 6 months infants can perceive universal phonemic distinctions, but in the second half of the first year they become more attuned to their own language and lose some of the distinctions in other languages. See, for example, Vihman (1996).

The work described here addresses a developmental stage analogous to the following period. Phonemes have been acquired, syllables of the infant's own language can be produced, then patterns of the teacher's syllabic output are acquired and word forms found in these patterns are learnt (without their semantics). Initial babbling becomes biased towards the teacher's speech, leading to word form learning.

Our work is based on the thesis that one-to-one dialogs between teacher and learner are essential for the acquisition of language. We take the term "dialog" in a broad sense to include early stages when the infant responds vocally to an adult utterance, but not with fully formed words. Research has shown that exposure to language on television does not compensate for lack of one-to-one interaction. For instance, the hearing children of deaf parents who cannot speak have severely impeded language skills until they mix with other speakers (Clark 2009, p. 39 ff).

3 Neuronal organization

Our approach in investigating word form acquisition separately from learning to use language meaningfully is also supported by neuroscientific research.

Dual processing streams

There is significant evidence that dual systems are needed for language processing, involving modular regions of activity as well as shared areas. On the one hand there is *implicit* learning of patterns and procedures, without intentional shared reference. On the other hand there is *explicit* declarative learning, in which there is joint attention between teacher and learner, and reference to objects, actions or relationships, sometimes described as item learning.

This dichotomy is also described as a dorsal pathway concerned with sub-lexical processing, object interactions and phonetic decoding, in contrast to a ventral pathway specialising in object identification and whole word recognition. This functional segregation is also characterised as a motor-articulatory system on the one hand and a conceptual system on the other (Hickok and Poeppel 2004; Saur and others 2008). We have adopted this dichotomy in a simplistic manner in the implementation of a language learning robot.

The role of mirror neurons in speech processing

In our experiments perception and production of speech are based on simulated mirror neuron type structures: the same elements reflect components of perceived speech and generate synthesized speech. The mirror neuron thesis has limi-

Table 1: The CMU phoneme set. (The phoneme *zh* does not occur in our corpora)

aa	odd	k	key
ae	at	l	lee
ah	hut	m	me
ao	ought	n	knee
aw	cow	ng	ping
ay	hide	ow	oat
b	be	oy	toy
ch	cheese	p	pee
d	dee	r	read
dh	thee	s	sea
eh	Ed	sh	she
er	hurt	t	tea
ey	ate	th	theta
f	fee	uh	hood
g	green	uw	two
hh	he	v	vee
ih	it	w	we
iy	eat	y	yield
jh	gee	z	zee

tations: summarising recent research Hickok (Hickok 2010) points out that speech perception can be achieved without motor speech ability. However, in specific situations "there must be a mechanism for the interaction of sensory and motor systems during perception of speech" (ibid).

The acquisition of preliminary word form recognition abilities illustrates both these points. Firstly, there is the well known asymmetry in language acquisition, in that infants can perceive more than they can express, due to immature articulatory development, (Vihman, dePaolis, and Keren-Portnoy 2009) and (Lyon et al. 2009).

But then, secondly, consider the learning process, as the infant tries out pre-lexical productions in babbling, and slowly converges on those that match components in the speech of the carer. Oudeyer (Oudeyer 2006, p. 148) asks why monkeys and apes do not have a speech system like humans and concludes that "it is babbling that makes the difference". He notes that human infants, and his own synthetic agents, "spontaneously try many articulatory configurations and systematically repeat these attempts. In other words they 'train themselves' ". See also (Pulvermuller 2002, p. 51).

When we consider the universal human ability to learn to talk it is essential to integrate perceptive and productive skills, supporting a form of mirror neuron theory applied to a specific situation.

4 Experimental work

Overview

These experiments first aim to show how words begin to be recognized without meaning attached. We make the following assumptions:

- Communicative ability is learnt through interaction with a teacher.
- Lesa has the intention to communicate and therefore can react positively to reinforcement.
- It acts autonomously, but is influenced by what it "hears".
- It practices turn taking in a dialog.

The scenario is a teacher in dialog with a simulated prelinguistic robot, analogous to a human infant aged about 6 to 14 months. The teacher's input is taken from the recorded speech of participants who were engaged to converse with the robot Kaspar. They were asked to teach it the names of shapes, treating it as a young child, and to speak as came naturally, but were given no further directions. However, an analysis of their input shows that it has the characteristics of child directed speech (CDS). Further details are in (Saunders, Nehaniv, and Lyon 2010). This contrasts with other examples of robot directed speech: it seems necessary to have a humanoid robot like a small child to evoke appropriate behaviour in the teacher

The teachers' transcribed speech was collected in a corpus and used in these experiments. Dataset 1 consists of 3276 words; Dataset 2, which includes Dataset 1, has 8148 words. An example of the teacher's speech is shown in Table 2.

As perceived by Lesa the teacher's speech is a string of phonemes, not yet segmented into words. Initially the infant's response is a string of random syllables. However its output becomes biased towards the syllables "heard" from the teacher, and eventually it will by chance produce a single syllable word "recognized" by the teacher. Lesa is then "rewarded" for producing this word which is latched in its memory with start and end of word markers, a candidate for future productions along with other syllables. From the start a dialog is set up, in which the teacher speaks to the infant, who responds.

Phonemes are represented by the CMU set, using 38 of 39 (CMU 2008). This consists of 15 vowels and 23 consonants, (Table 1). The software used to produce them comes from SysMedia (SysMedia 2009). The phonemic transcription of speech is taken from an intermediate stage in a speech recognition process. Note there can be a fuzzy match: for instance "round" is represented as "r-ah-n-d", "r-aw-n-d" (see Table 2) and "r-ae-nd" after consonant clustering (see Table 7).

Syllables as primary elements of communication

Infants can recognize phonemes, as discussed above in the comments on categorical perception. They can produce vowel sounds such as "ah", "ur" "oo". From an early age they can distinguish consonants in minimal pairs of words such as "cat" and "rat". However, phoneme consonants cannot usually be expressed in isolation, apart from a few exceptions such as "shh". The most primitive vocal utterances are syllabic (Vihman, dePaolis, and Keren-Portnoy 2009; Clark 2009; de Boisson-Bardies 1999). In our simulation the infant will perceive phonemes and process them into produceable syllables.

There are 4 types of syllables: V, CV, CVC and VC, where V is a vowel and C is one or more consonants. In our notation C can represent either a single phoneme or a phoneme cluster so "square" (s-k-w-eh-r) and "box" (b-aa-k-s) are both of the form CVC (rather than $C^{+}VC^{+}$). Clus-

Table 2: An example of the speech from a teacher to Lesa, with its phonemic transcription

can you see thats a star its a picture of a star there yeah can you see the round there we turn it round and we have can you see thats a picture of a heart
k ah n y uw s iy th ae t s ah s t aa r ih t s ah p ih k ch er ah v ah s t aa r dh eh r y ae k ah n y uw s iy dh eh r ah n d dh eh r w iy t er n ih t r aw n d ae n d w iy hh ae v k ae n y uw s iy th ae t s ah p ih k ch er ah v ah hh aa r t

Table 3: Words in *lexset* which the teacher is helping Lesa to learn in these experiments

Words in *lexset*: **box, heart, moon, round, sun, square, star**
Phonemically represented, after consonant clustering: **b-aa-ks, hh-aa-rt, m-uw-n, r-ae-nd** **s-ah-n, skw-eh-r, st-aa-r**

tering consonants into allowable strings is a first step in processing. We extract from the SCRIBE corpus almost all possible consonant clusters that occur in English (SCRIBE 2004).

Early babbling in infants is observed to be single syllable, typically of the form CV repeated as in "da da da" (Clark 2009, chapter 5). Production of CVC forms comes later. Infants have a restricted repertoire since their articulatory mechanisms are immature. However, neither the real robot Kaspar nor its synthetic counterpart Lesa have articulatory development modelled so this restriction is not implemented.

Program design

The program models developmental stages in the following way. Contact the first author for the code.

- Lesa produces random babble, any of the 4 types of syllable can be produced with combinations of consonants and vowels. Any allowable consonant cluster may be produced. The syllable type is chosen at random, and there is a 50% chance that syllables of type V, CV and VC will be repeated. An example is given in Table 4, obviously as it is random it will vary each time the program is run. The repetition is purely cosmetic at present, but may be used later when we investigate the role of rhythmic motor skills (Vihman, dePaolis, and Keren-Portnoy 2009).
- The teacher speaks to Lesa, see Table 2, encouraging it to learn the names of different shapes. The speech is represented as a stream of phonemes. In Experiment 1 the input from Dataset 1 consists of 3276 words, 10301 phonemes; in Experiment 2, from Dataset 2, it consists of 8148 words, 25551 phonemes. These corpora were collected as described above.
- Lesa starts with a store of syllable types in its memory, analogous to the sounds produced in ambient speech. It perceives the phoneme strings from its teacher, and each time a syllable is perceived its frequency is augmented. Syllables may overlap, so (using letters as peudo-phonemes) *a s t a r* would generate syllables *a, a st, st a, st a r, a r*.
- At the next stage Lesa's output is still quasi-random, but biased towards more frequently perceived syllables. Initially, the random selector picks one of the stored syllable types, all equally likely. Now frequently occurring syllables have a greater chance of being picked. For instance, if there were 4 syllables, *s1, s2, s3, s4* initially each would have 1/4 chances of being produced. If *s2* occurs 2 times, *s3* occurs 4 times and the others do not occur at all, then *s1* will have 1/10 chances of being produced, *s2* will have 3/10 chances, *s3* will have 5/10 chances and *s4* 1/10.
- The teacher has a lexicon, *lexset*, containing the words s/he wants to teach. In this simulation, for instance, *lexset* contains the words box, heart, moon, round, sun, square, star. See Table 3. (This is an over-simplification since the teachers actually used a variety of words to describe the shapes, e.g "crescent" and "smile" for "moon".)
- The simulated infant also has a lexicon, empty to start with.
- By chance Lesa will eventually produce a word
- The teacher will "reward" Lesa, by marking this event, and the word will then be entered in Lesa's lexicon.
- Next time the random selector is picking a syllable type the word in the lexicon will be a candidate to be chosen, alongside the syllable types

The "reward" is purely metaphorical currently. However, when we carry out similar experiments with a human talking to a real robot we hope to identify the teacher's approval - by words and by prosodic markers

Table 7 shows how many words are recognized, on average, as Lesa goes through a given number of utterances.

5 Discussion

Each time the simulation is run a different result is produced, because of the role of random selectors. However, though

Table 4: Example of initial random babbling. For syllables of the form V, CV and VC there is a 50% chance of repetition.

```
(’ao’, ’ks’) (’ao’, ’ks’) (’ao’, ’ks’)
(’ch’, ’ah’, ’ld’)
’ih’
(’y’, ’ah’, ’p’)
’ey’ ’ey’
(’uh’, ’dh’) (’uh’, ’dh’)
(’th’, ’iy’)
(’skw’, ’ey’)
(’ae’, ’m’)
```

Table 5: Example of babbling biased towards teacher’s speech; (’m’, ’uw’) is part of “moon”, (’s’, ’iy’, ’dh’) is part of “see this”.

```
(’w’, ’ah’) (’w’, ’ah’) (’w’, ’ah’)(’s’, ’iy’, ’dh’)
(’uw’, ’s’) (’uw’, ’s’)
(’t’, ’ah’, ’b’)
(’m’, ’uw’)
(’d’, ’ah’)
(’aa’, ’t’) (’aa’, ’t’) (’aa’, ’t’)
(’t’, ’ay’, ’m’)
(’s’, ’iy’, ’dh’)
’aa’ ’aa’
```

the results vary they all show Lesa producing one or more words after 20 utterances.

There is a higher frequency of occurrence of words to be taught, in the larger dataset, however, it does not result in the production of more words by Lesa (Table 7). The larger dataset generates a larger number of CVC syllables, thus reducing the chance that the random selector will pick any particular one. Dataset 1 produces 511 CVC syllables, Dataset 2 produces 760.

Once a word has been produced and stored in *ilex*, the infant’s lexicon, it then becomes a candidate for selection alongside the 4 syllable types. Thus after the first word is found there is a 1 in 5 chance of its being produced again. This means that early words are produced rather frequently, as observed with human infants.

However, in our simulation, as more words are found and stored the occurrence of other syllables consequently declines. The rate of word discovery attenuates (Table 7). These results depend on the parameters of the method currently in use, which could be adjusted.

6 Conclusion

This paper focuses on the transition from babbling to the recognition of first word forms. Our simulation shows how this transition might occur in a synthetically enabled linguistic agent or robot, through dialog with a human teacher. One reason this matters is that it is a necessary step to recognize word forms before words are associated with their semantics.

Table 6: Example of Lesa’s output with teacher “rewarding” production of wanted words

```
Utterance number 20

(’kw’, ’eh’, ’r’)
’ah’ ’ah’ ’ah’ ’ah’ ’ah’ ’ah’
’uw’

s-ah-n
word produced

’ah’

infant lexicon contains 4 words
’sh-ey-p’, ’skw-eh-r’, ’m-uw-n’, ’s-ah-n’
```

Table 7: Statistical data on Experiment 1, using Dataset 1 and Experiment 2, using Dataset 2. *ilex* is the infant’s lexicon, initially empty

Expt1		Words in *ilex*
		After 20 utterances
Run 1	2	skw-eh-r, s-ah-n
Run 2	3	hh-aa-rt, sh-ey-p, r-ae-nd
Run 3	4	skw-eh-r, sh-ey-p, hh-aa-rt, m-uw-n
Run 4	2	s-ah-n, sh-ey-p
		After 50 utterances
Run 1	5	skw-eh-r, s-ah-n, sh-ey-p, st-aa-r, m-uw-n
Run 2	5	hh-aa-rt, sh-ey-p, r-ae-nd, skw-eh-r m-uw-n
Run 3	6	skw-eh-r, sh-ey-p, hh-aa-rt, m-uw-n, s-ah-n, r-ae-nd
Run 4	5	s-ah-n, sh-ey-p, m-uw-n, hh-aa-rt, skw-eh-r
Expt2		Words in *ilex*
		After 20 utterances
Run 1	3	s-ah-n, hh-aa-rt, skw-eh-r
Run 2	4	s-ah-n, skw-eh-r, st-aa-r, sh-ey-p
Run 3	2	skw-eh-r, sh-ey-p
Run 4	2	skw-eh-r, sh-ey-p
		After 50 utterances
Run 1	4	s-ah-n, hh-aa-rt, skw-eh-r, m-uw-n
Run 2	6	s-ah-n, skw-eh-r, st-aa-r, sh-ey-p, hh-aa-rt, m-uw-n
Run 3	4	skw-eh-r, sh-ey-p, m-uw-n, hh-aa-rt
Run 4	3	skw-eh-r, sh-ey-p, hh-aa-rt

If the robot perceives a stream of sound it needs to extract salient word forms to which meaning can be attached. It is necessary to segment the speech stream, which the earlier process of word form detection can aid. There is evidence that the ability to segment involves the integration of prosodic information, phonotactic constraints, facial expression together with the recognition of repeated word forms and holophrases that act as anchor points in an utterance. It can be shown that it is easier for the perceiver to decode speech if it is segmented appropriately, when the entropy of the stream of sound declines (Lyon, Nehaniv, and Dickerson 2007).

In practice the development of the ability to segment a speech stream into pre-referential words and phrases overlaps with the acquisition of semantic understanding and the recognition of primary language structure. However, we initially investigate these processes separately in order to understand each strand better.

In our simulation we have extracted from observed infant development those factors that seem relevant to language acquisition by a robot. Thus we have not modelled the articulatory constraints that are typical of infant productions. However, our approach is in accord with the dual processing system observed in recent neuroscientific research, and results in the desired effect that the agent learns to produce word forms. It shows how continuity between babbling patterns and early word production can be modelled.

Following the work described here we have implemented a real time version. The teacher's speech to Lesa is converted to a stream of phonemes, processed by Microsoft SAPI 5.4. Lesa's output is processed by the eSpeak speech synthesizer. This is being ported to the real iCub robot. We are working on automatically capturing positive feedback through prosodic information, and will see whether this corresponds to identifiable verbal expressions.

Acknowledgements This work is supported by the European Commission under project grant FP7-214668 for ITALK : Integration and Transfer of Action and Language Knowledge in Robots.

References

Bloom, P. 2002. *How Children Learn the Meaning of Words*. MIT Press.

Clark, E. V. 2009. *First Language Acquisition*. Cambridge,UK: Cambridge University Press, 2nd edition.

CMU. 2008. The CMU pronouncing dictionary. www.speech.cs.cmu.edu/cgi-bin/cmudict. [visited June 2010].

Dautenhahn, K.; Nehaniv, C. L.; Walters, M. L.; Robins, B.; Kose-Bagci, H.; Mirza, N. A.; and Blow, M. 2009. Kaspar - a minimally expressive humanoid robot for human-robot interaction research. *Applied Bionics and Biomechanics, Special Issue on 'Humanoid Robots'* 6(3):369–397.

de Boisson-Bardies, B. 1999. *How Language Comes to Children*. MIT Press.

Hickok, G., and Poeppel, D. 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92.

Hickok, G. 2010. The role of mirror neurons in speech and language processing. *Brain and Language* 112. Preface to Special Issue.

Lyon, C.; Sato, Y.; Saunders, J.; and Nehaniv, C. L. 2009. What is needed for a robot to acquire grammar? some underlying primitive mechanisms for the synthesis of linguistic ability. *IEEE Transactions on Autonomous Mental Development* 1:187–195.

Lyon, C.; Nehaniv, C. L.; and Dickerson, B. 2007. Clues from information theory indicating a phased emergence of grammar. In Lyon, C.; Nehaniv, C. L.; and Cangelosi, A., eds., *Emergence of Communication and Language*. Springer.

Morgan, J., and Demuth, K. 1996. Signal to syntax: an overview. In Morgan, J., and Demuth, K., eds., *Signal to Syntax*. Lawrence Erlbaum.

Oudeyer, P.-Y. 2006. *Self-organization in the Evolution of Speech*. Oxford University Press.

Pulvermuller, F. 2002. *The Neuroscience of Language*. Cambridge University Press.

Roy, D., and Pentland, A. 2002. Learning words from sights and sounds: A computational model. *Cognitive Science* 26:113–146.

Saunders, J.; Lyon, C.; Förster, F.; Nehaniv, C. L.; and Dautenhahn, K. 2009. Constructivist approach to robot language learning via simulated babbling and holophrase extraction. In *Proc. 2nd International IEEE Symposium on Artificial Life (IEEE Alife 2009), Nashville, Tennessee, USA*. IEEE.

Saunders, J.; Nehaniv, C. L.; and Lyon, C. 2010. The acquisition of word meanings by a humanoid robot through interaction with a tutor. In Dautenhahn, K., and Saunders, J., eds., *New Frontiers in Human-Robot Interaction*. John Benjamin. Forthcoming.

Saur, D., et al. 2008. Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences* 105(46).

SCRIBE. 2004. Spoken Corpus Recordings In British English. www.phon.ucl.ac.uk/resource/scribe/scribe-manual.htm. [visited May 2010].

SysMedia. 2009. Sysmedia word and phoneme alignment software. www.sysmedia.com.

Tomasello, M. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Vihman, M. M.; dePaolis, R. A.; and Keren-Portnoy, T. 2009. A dynamic systems approach to babbling and words. In Bavin, E. L., ed., *The Cambridge Handbook of Child Language*. CUP. 163–182.

Vihman, M. M. 1996. *Phonological Development: the Origins of Language in the Child*. Blackwell.