A multimodal corpus of speech to infant and adult listeners

Elizabeth K. Johnson
University of Toronto Mississauga, 3359 Mississauga Road N.,
Mississauga, ON L5L 1C6, Canada (elizabeth.johnson@utoronto.ca)

Mybeth Lahey
Max Planck Institute for Psycholinguistics, POB 310, 6500 AH Nijmegen, The
Netherlands (mybeth.lahey@mpi.nl)

Mirjam Ernestus
Centre for Language Studies, Radboud University Nijmegen, Postbus 9103,
6500 HD Nijmegen, The Netherlands (m.ernestus@let.ru.nl)

and Anne Cutler
The MARCS Institute, University of Western Sydney, Penrith South, NSW 2751,
Australia (A.Cutler@uws.edu.au)

Running Title: A corpus of speech to infant listeners

Corresponding author: Elizabeth. K. Johnson, University of Toronto Mississauga, 3359
Mississauga Road N., Mississauga, ON L5L 1C6, Canada
Abstract

An audio and video corpus of speech addressed to 28 11-month-olds is described. The corpus allows comparisons between adult speech directed towards infants, familiar adults and unfamiliar adult addressees, as well as of caregivers’ word teaching strategies across word classes. Summary data show that infant-directed speech differed more from speech to unfamiliar than familiar adults; that word teaching strategies for nominals versus verbs and adjectives differed; that mothers mostly addressed infants with multi-word utterances; and that infants’ vocabulary size was unrelated to speech rate, but correlated positively with predominance of continuous caregiver speech (not of isolated words) in the input.

PACS numbers: 43.71.Ft

Keywords: infant directed speech, word learning, vocabulary development, language acquisition, spoken language corpus
1. Introduction

Among the many remarkable aspects of language acquisition is the ease and rapidity with which young children acquire a vocabulary. Vocabulary learning begins well before its results can be turned into useful conversations; the earliest produced words appear generally very late in the first year of life, but evidence for recognition of spoken forms, including the child’s own name [1], names of body parts [2, 3], and the names of caregivers [4], has been found many months earlier.

The language input that infants are exposed to is, of course, crucial in this acquisition process. Although some have argued that the production of isolated words plays a crucial role in early language development [5, 6], the available evidence overwhelmingly suggests that speech directed to infants consists largely of multi-word utterances [7, 8], forcing young language learners to extract word forms from surrounding speech context in order to store them in memory for subsequent recognition. A lot is known about the characteristics of input to infants in the first year of life. For example, in comparison with speech directed to adult listeners, speech to infant listeners generally has a slower speaking rate, exaggerated prosody, shorter utterances, greater expressive affect, and an expanded vowel space [e.g., 8, 9, 10]. Importantly, many of these properties have been shown to correlate with infant success at word-form perception and learning [11, 12, 13, 14], crucially linking the nature of the language input to subsequent language learning success.

Laboratory studies of infant-directed speech, from which this picture has arisen, have typically contrasted productions by a mother addressing her own infant versus the same mother addressing an unfamiliar adult (often the experimenter). This approach has
enabled researchers to collect well-controlled clear audio recordings of infant- and adult-directed speech. However, the speech the mother directs to an unfamiliar experimenter encountered in a laboratory setting may not be representative of the speech the mother would direct to other more familiar adults commonly encountered in the child’s everyday environment. In addition the child’s typical environment may also often include multiple simultaneous interlocutors, forcing the child to separate overlapping speech streams in order to extract useful linguistic information from spoken input. Some of the properties of infant-directed speech mentioned above have also been studied in more naturalistic recordings made in the home, whereby most of these naturalistic recordings with children under a year have involved dense recordings with a single child [e.g., 8]. Such dense corpora are highly useful, but do not readily allow researchers to assess generalizability or to examine the relationship between different types of caregiver input and children’s developing language skills. That is, existing corpora tend to be naturalistic but less generalizable, or generalizable but less naturalistic.

In this report we describe an extensive new corpus that (a) involves recordings of input provided (to 28 Dutch-learning infants) at the very onset of the natural word-teaching period, namely just before the first birthday; (b) involves speech from multiple caregivers interacting with the infant together, as well as speech between caregiver(s) and experimenter; (c) includes a video record from two separate angles to avoid loss of relevant data; and (d) also compares teaching of words from different word classes (noun, proper name, verb, adjective).

Data from this corpus will be generalizable, but also naturalistic. Such an extensive corpus consisting of both semi-structured word teaching activities and free
interactions allows a wide variety of differing questions to be addressed, including, but not limited to, questions about the acoustics of the speech input, the relation of speech characteristics to eye contact, potential differences in adult-to-child interactions during explicit word teaching tasks versus free interaction, the consistency of speech acoustics across talkers and across interlocutor groups, and the variation of speech acoustics as a function of word class. The primary purpose of this summary overview is to announce the public availability of these data to the field of infant speech perception research. In addition we show how the data can help to answer three specific questions of general interest to infant speech perception researchers: how does mean length of utterance vary across speech directed to infants, to familiar adults and to unfamiliar adults? Are words of different word classes taught in the same way? Have children with larger productive vocabularies been exposed to more isolated word tokens? Answers to these questions, and to many more which the corpus will enable researchers to address, will help to improve our understanding of the crucial relationship between speech input and language learning.

2. Data Acquisition

The corpus comprises 33 hours of spoken language interaction in 65 play sessions recorded between September 2005 and August 2007 at the Max Planck Institute Baby Research Centre in Nijmegen, The Netherlands. Each recording involved an infant, the infant's mother or father, and an experimenter, with most recordings also involving an additional caregiver. Both free play and word teaching activities were recorded, as well as adult-to-adult conversations. All interactions were audio-taped using radio-controlled
microphones clipped to the two caregivers, and videotaped from perpendicular angles (See Figure 1). To assess vocabulary development, the Dutch version of the MacArthur Bates Child Developmental Inventory (N-CDI, short form [15]) was collected at the time of the first session, and a follow-up N-CDI was collected approximately one year later.

2.1. Participants

Twenty-eight triads were recorded; each triad included a typically-developing infant with reportedly normal hearing (20 males; 8 females), a parent of the infant, and an additional primary caregiver (usually the other parent, or grandmother). All participants lived in the Nijmegen region at the time of the recording sessions and all but one were native Dutch speakers. Parental region of origin, education level, and approximate age were noted.

Infant participants were approximately 11 months old at the time of the first recording session (Range = 11:0 to 12:6; M= 351 days; SD=11:7). Half of the participants (N=14) were recorded on two further occasions within 20 days after the first session, but not all participants were able to commit to three sessions. For nine participants the corpus contains only one session, and for five participants two sessions.

2.2. Recording Sessions

Sessions were recorded in a quiet room at the Baby Research Centre in Nijmegen, and ranged in duration from 21 to 40 minutes (M= 30.5 minutes). Recording of each session began with a five-minute free play period, followed by 12 minutes of word teaching. All recording sessions ended after various activities designed to elicit unscripted adult-to-
adult interactions. Thus, transcribed interactions involved ample free interactions and discussions in addition to the targeted word teaching activities.

In the first session, caregivers were given four words to teach their child: a proper name (Zanthe, Mirre, Roemer, or Tigo), a common noun (cactus ‘cactus’, koffer ‘suitcase’, zadel ‘saddle’, or masker ‘mask’), an adjective (glanzend ‘shiny’, krullend ‘curly’, puntig ‘pointy’, or harig ‘hairy’), and a verb (stampen ‘to stomp’, krabben ‘to scratch’, buigen ‘to bow’, or knikken ‘to nod’). Each triad was assigned one of each of the four word types to teach their child, with different triads given different combinations of words. All these bisyllabic target words had a trochaic stress pattern and were deemed unlikely to be known by the average 11-month-old. To facilitate word teaching, props were provided, including a doll with the to-be-taught name, and relevant objects (e.g. a toy cactus, or a shiny colander). If caregivers returned for a second session, they were given four new words to teach their child. For example, if a child learned cactus, Tigo, glanzend, and stampen on the first visit, they might be taught koffer, Mirre, harig, and krabben on the second visit. Immediately following each Day 1 or 2 word teaching task (and before the recording ended), caregivers spent some time filling out paperwork together. Much of the read speech in this corpus resulted from the caregivers reading these forms to each other. To elicit adult-to-adult conversation, the caregivers were encouraged to work together on answering the questions on the forms. Spontaneous conversations between the adults also occurred during the recording sessions. The experimenter returned towards the end of the session to chat with the caregivers. Recorded interaction with the experimenter included free conversation as well as discussion of paperwork and word teaching strategies. The first two visits always
involved two caregivers plus an experimenter; the third visit involved only one caregiver. In the third visit, the caregiver reviewed the words taught during the first two sessions.

At the outset of all sessions, caregivers were asked to spend equal time teaching each word type (noun, verb, adjective, name). They were not told how to teach the words (e.g., they were not told to put target words in complete sentences or to avoid diminutive forms); they were however told that their children would be tested later on their knowledge of the words. First and second visits thus finished with a Headturn Preference Experiment [16] to assess infants’ recognition of the four word forms taught during that particular visit. The third visit finished with a Preferential Looking Experiment [16] to test comprehension of the eight words taught over the course of all three sessions.

2.3. Vocabulary Measures

Parents of all 28 participants filled out the Dutch version of the MacArthur-Bates Words and Sentences Child Development Inventory (N-CDI) at the time of the first recording sessions. For 20 of the 28 participants, an additional follow-up N-CDI was filled out near the child’s second birthday. (Comparison to standardized N-CDI word production scores placed five of these 20 in the top quartile, and four in the bottom quartile. Thus, by age two, our participants appeared representative of the Dutch-learning toddler population.)

2.4. Transcription

The corpus includes over 57,000 utterances, orthographically transcribed using the software program ELAN (http://tla.mpi.nl/tools/tla-tools/elan/; [17]). Utterances were defined as breath groups, and were marked as IDS (infant-directed speech), FADS
(speech addressed to a familiar adult), UADS (speech addressed to an unfamiliar adult), or RS (read speech). Assignment of utterances to register type was based on reference to both audio and video recordings. Each adult’s productions were coded on a separate text tier labeled Caregiver 1, Caregiver 2, or Experimenter. Utterance onset and offset was marked, to enable subsequent analyses of speech overlap and turn structure. Instances where transcribers could not identify the words being spoken by the adults were coded as “XX”. No attempt was made to transcribe the children’s productions.

INSERT FIGURE 1 AND TABLE 1 ABOUT HERE

2.5. Summary Data

The corpus contains 29,316 utterances directed to the infant (i.e., IDS), 9,834 utterances to a familiar adult (i.e., FADS, or speech between caregivers), 16,754 utterances to an unfamiliar adult (i.e. UADS, or speech between caregiver and experimenter), and 1,817 read utterances. The analyses reported below are based on all spontaneous productions made during the entire approximately 30 minute recording sessions (excluding the read speech). Caregivers’ utterances frequently overlapped, even when the infant was being addressed; on average, 27% of the IDS utterances infants heard overlapped with another utterance. Variability in this measure was large, however, with some infants hearing as little as 12% and some hearing as much as 46% overlapping IDS utterances.

Although caregivers did not avoid speaking at the same time to the child, they did clearly modify their speaking style when addressing their child. The effect of speech register (i.e., speaking style used when addressing infants, a familiar adult, versus an unfamiliar adult) on utterance length is illustrated in Figure 1 (Panel A). As predicted, utterances were shorter in IDS than in either adult-to-adult speech register. Interestingly,
we also found that utterance length differences tended to be greater between UADS and IDS than between FADS and IDS. Speech rate differences between the three registers exhibited a different pattern: although IDS differed from both UADS and FADS, UADS and FADS speech rates were very similar (see Table 1). Although several studies have examined how IDS differs from FADS or UADS, this is the first report of which we are aware to treat UADS, FADS, and IDS as three distinct registers. Our overall analyses suggest that studies using only UADS to represent all ADS (as in laboratory studies comparing speech between a mother and child and a mother and experimenter) may lead some IDS/ADS differences to appear larger than in studies comparing IDS to all forms of adult directed speech (including conversations between parents).

Caregivers produced many target tokens (Figure 1, Panel B). Inflected forms were produced more often for adjectives and verbs than for nouns and proper names. Despite the instructions to focus equally on training each word type, caregivers produced fewer tokens of adjectives than nouns, names, or verbs. Given that children’s early productions contain few adjectives [18], caregivers may have strategically focused more energy on the word types they thought they had the best chance to successfully teach to their child.

Finally, we examined maternal IDS in relation to individual vocabulary development. Mothers’ IDS rates varied greatly (M = 4.1 syllables per second; Range = 3.0 to 5.6). Slower speech rates facilitate young children’s word recognition [14], and we therefore predicted a relation of individual mothers’ IDS rate to their children’s N-CDI score at 11 months. However, we observed no such relationship of mother’s speech rate modification (FADS-IDS difference) with children’s measured vocabulary size (see Figure 1, Panel C). Mothers also varied greatly in the proportion of tokens produced in
isolation (M=28%; Range = 13% to 46%; note that we included fillers and vocatives in our count, so that this number is higher than in some past studies that excluded such forms [e.g., 8]). Prior studies have suggested that exposure to isolated words promotes language development [5, whose count also included all isolated tokens], and we thus predicted that caregivers' production of few isolated words might be linked to smaller productive vocabulary size. However, after excluding the data from one mother who only produced a single token of a trained word, we found the opposite pattern: the production of a greater proportion of multi-word utterances relative to single-word utterances was associated with infants having larger productive vocabularies at 11 months of age, r(26) = .52, p = .005. There are at least two plausible explanations for this finding: 1) mothers modify their speaking style to suit their infants’ level of linguistic skill (e.g., [19]), or 2) exposure to multi-word utterances promotes (rather than hinders) vocabulary development in young infants.

3. Conclusion

Children’s vocabularies expand at an astonishing rate in the second year of life. This corpus focuses on the language input children are receiving when they are on the brink of this expansion [18]. The corpus is specifically designed to allow researchers to address key questions that are difficult or impossible to address in other corpora examining language input at this crucial stage of development. The findings summarized in this initial report already motivate more nuanced definitions of ‘adult-directed’ versus ‘infant-directed’ speech, and provide new insights into the role of early input in word discovery.
Acknowledgements

Corpus construction was funded by NWO-SPINOZA ("Native and nonnative listening"; A. Cutler). The second, third and fourth authors are affiliated to both the Max Planck Institute of Psycholinguistics and the Radboud University Nijmegen. We thank Tijl Grootswagers for assistance in programming and Angela Khadar for assistance in data acquisition.
References


Table 1. Average speech rate in syllables per second, by register, caregiver type and recording session. Only data for the first two sessions are reported, since on the third visit no second familiar adult was present and no grandmothers came in.

<table>
<thead>
<tr>
<th></th>
<th>Mother</th>
<th></th>
<th>Father</th>
<th></th>
<th>Grandmother</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Session 1</td>
<td>Session 2</td>
<td>Session 1</td>
<td>Session 2</td>
<td>Session 1</td>
</tr>
<tr>
<td>IDS</td>
<td>4.08</td>
<td>4.02</td>
<td>3.89</td>
<td>4.09</td>
<td>3.88</td>
</tr>
<tr>
<td>FADS</td>
<td>4.73</td>
<td>4.82</td>
<td>4.75</td>
<td>5.25</td>
<td>4.87</td>
</tr>
<tr>
<td>UADS</td>
<td>4.70</td>
<td>4.81</td>
<td>4.93</td>
<td>5.32</td>
<td>4.84</td>
</tr>
</tbody>
</table>
Figure Caption

Figure One: A) Average utterance length in words, as a function of speech register (infant-directed speech, familiar-adult-directed speech, and unfamiliar-adult-directed speech). Error bars indicate standard error. B) Total number of infant-directed target words produced by caregivers. Base forms are indicated in light grey (e.g., ‘knikken’ for ‘knikken’), inflected forms are indicated in dark grey and included comparatives (e.g. ‘hariger’ for adjective ‘harig’), diminutives (‘koffertje’ for noun ‘koffer’), conjugations (e.g., ‘stamp’ for verb ‘stampen’), possessives (e.g., ‘Roemers’ for ‘proper name ‘Roemer’), etc. C) Mother’s degree of speech rate modification plotted as a function of children’s productive vocabulary scores at the time of the recording. D) Example screenshot of two simultaneous camera angles recorded during each session. In general, the gaze direction of all participants can be retrieved from the video record.