GLU 2017 International Workshop on Grounding Language Understanding
25 August 2017, Stockholm, Sweden

Building a multimodal lexicon: Lessons from infants' learning of body part words

Rana Abu-Zhaya[1]*, Amanda Seidl*[1]*, Ruth Tincoff*[2]*, and Alejandrina Cristia*[3]

[1]Speech, Language and Hearing Sciences, Purdue University, West Lafayette, IN, USA
[2]The College of Idaho, Caldwell, USA
[3]LSCP, Département d' Études Cognitives, ENS, EHESS, CNRS, PSL Research University, Paris, France

rabuzhay@purdue.edu, aseidl@purdue.edu, rtincoff@collegeofidaho.edu, alejandrina.cristia@ens.fr

Abstract

Human children outperform artificial learners because the former quickly acquire a multimodal, syntactically informed, and ever-growing lexicon with little evidence. Most of this lexicon is unlabelled and processed with unsupervised mechanisms, leading to robust and generalizable knowledge. In this paper, we summarize results related to 4-month-olds' learning of body part words. In addition to providing direct experimental evidence on some of the Workshop's assumptions, we suggest several avenues of research that may be useful to those developing and testing artificial learners. A first set of studies using a controlled laboratory learning paradigm shows that human infants learn better from tactile-speech than visual-speech co-occurrences, suggesting that the signal/modality should be considered when designing and exploiting multimodal learning tasks. A series of observational studies document the ways in which parents naturally structure the multimodal information they provide for infants, which probably happens in lexically specific ways. Finally, our results suggest that 4-month-olds can pick up on co-occurrences between words and specific touch locations (a prerequisite of learning an association between a body part word and the referent on the child's own body) after very brief exposures, which we interpret as most compatible with unsupervised predictive models of learning.

Index Terms: Natural language acquisition, mapping knowledge in the world, grounded dialogue, supervision, reinforcement, human infancy, body part words

1. Introduction

One of the tasks humans excel at is acquiring a multimodal, syntactically informed, and ever-growing lexicon with fairly little evidence from unlabeled and unprocessed input in environments that are more likely to be cluttered than not. In contrast, artificial learners do not fare as well, even in limited tasks and when provided with much more evidence and labeling in simple object identification scenes. For example, the speech recognizer described in [1] requires supervised training with 350 million words to achieve performance comparable to first-pass transcribers. This is more than twice the amount of experience provided to American children between birth and 10 years of age, and about 40 times as much input provided to children in a hunter-farmer community in the Bolivian Amazon [2]. Humans also process the input they receive in an unsupervised manner: direct feedback on whether parsing is being made correctly or incorrectly is extremely rare, particularly at early stages of processing. Most saliently, caregivers are not aware of (and therefore cannot correct) word segmentation errors in the child's first year of life. Despite the lower quantity of data, the great majority of which is unlabeled, and the lack of effective external feedback, human children not only learn to parse the spoken signal at a level at the very least comparable to Xiong's state-of-the-art parser, but they also learn a great deal more. They learn a lexicon that contains multimodal meaning and they distinguish nuanced ways in which words can be used in syntactic contexts allowing them to make grammaticality decisions on word sequences they have never encountered. In this paper, we summarize evidence on how infants come to learn body part words by around 6 months of age. In addition to providing direct experimental evidence on some of the Workshop assumptions, we suggest several avenues of research that may be useful to those developing and testing artificial learners.

2. Not all multimodal information is created equal

Most artificial learners are trained solely on visual and auditory signals (though see an exception in [3]). The system in [4] for instance, relies on images of objects coupled with child-directed speech. Similarly, the systems in [5, 6, 7] learn using visual information such as the color and shape attributes of objects, and human gestures, critically, through extremely simple scenes. Yet, such a combination of visual and auditory cues alone, presented in an uncluttered, highly controlled environment, is not common in the input to human infants. Rare pathological cases aside, humans have frequent access to at the very least one more modality, namely touch, which we argue may be more important (at least in early stages of learning) than the visual modality. Specifically, in [8] we familiarized 4-month-olds with a continuous stream of syllables in a tactile or a visual condition. In the tactile condition, infants received a timed tactile stimulation of their elbow or knee that was always synchronous with a specific syllable sequence (e.g. *lepoga* was always timed with a touch to the knee). In the visual condition, infants received similar input, in the visual modality, by watching an experimenter touch her own eyebrow (or chin). Infants in both conditions were also touched, or observed a touch, on another location; this touch (e.g. on the infant's elbow) or visual cue (a touch by the experimenter on her own chin) was not consistently synchronous with a particular syllable sequence (e.g., *dobita* occurred once with a touch on the elbow and all other times without this touch; other touches to the elbow coincided with many other syllable sequences). After familiarization, infants were tested for their listening preferences for syllable se-

10.21437/GLU.2017-4

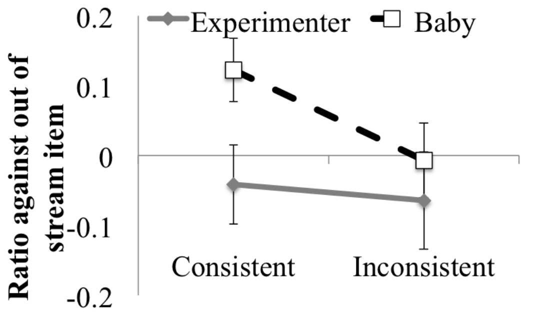

Figure 1: *Results from [8]: Infants looked longer to the out of stream item (a syllable sequence completely absent from the familiarization audio) than a syllable sequence present in the audio having received consistent touches only when touches were made on the infant's own body, and not the experimenter's. The inconsistent condition shows that this is not just a preference based on the presence or absence of a syllable sequence in the audio. The y-axis shows the ratio calculated as (LTx-LT_{OOS})/(LTx+LT_{OOS}), where LTx is looking times to the appropriate in-stream item (consistent or inconsistent), and LT_{OOS} looking times to the out of stream item.*

quences that had been paired with consistent versus inconsistent multimodal stimulation. The results in the tactile condition showed that infants' listening times to sequences that had been paired with consistent touches were shorter than listening times to sequences without consistent touches. In fact, the latter were treated the same as sequences that were not in the familiarization audio at all. No such differences were found in the visual condition, suggesting that infants showed no recognition of the sequences that were coupled with consistent visual cues alone. Thus, reliable self-focused speech+touch input led to learning, but similarly informative speech+visual input did not (see Figure 1). These results show how, in human infants, "language [is grounded] in the perceptual, emotional, and sensorimotor experience of the agent" - exactly as stated in the GLU Statement, and especially so for early word learning [9, 10, 11]. However, these results also suggest important asymmetries in coding of the different types of information. A model where all inputs are coded equally would predict similar learning in the visual and tactile conditions in this study - whereas we found that tactile signals, perhaps due to their social significance, are processed differently from visual signals. It may be worth pointing out that the rest of the contents of the lexicon also fail to strongly support a visual bias [12, 13, 14]. The other words 6-month-olds reportedly know are social words such as "mommy" and the child's name, and food items (such as "banana" and "cookie"). All of these are probably experienced in situations where the infant is highly aroused, for instance engaged in social play and feeding. These items likely involve proprioceptive/tactile cues and, in the case of food items, gustatory and olfactory cues. We will return to this in the "Future Directions" section.

3. Infants' dyadic partners provide lexically specific, highly structured, multimodal information

The next studies we summarize show that caregivers provide infants with information that may not be available to artificial learners today, but could be implemented in the future to aid model learners; we suggest that artificial learners trained on realistic input would benefit from extracting such structure. In two observational studies involving 4-month-old infants and their primary caregivers, we looked at whether speech+touch events comparable to those we manipulated in [8] may occur in infant-caregiver interaction [15, 16]. In [15] we provided caregivers with books depicting body parts, and books depicting animals (see Figure 2). In [16] we asked parents to teach made-up words for knee, elbow, and a control object (a vegetable brush; see Figure 3). In neither study did we say anything to caregivers about tactile stimulation; thus, caregivers were unaware of the fact that multimodality was our key research interest. Due to space limitations, we discuss results combining across studies. In both studies, bimodal touch+speech events, with the touch directly on the infant's body, were biased to body part discussion, occurring more frequently when caregivers talked about body part words than animal words [15] or when caregivers taught the body part words than the vegetable brush [16]. Moreover, caregivers systematically exaggerated cues in both streams of input creating speech+touch events that were physically different from speech-only and touch-only events: Touch events that were combined with speech were longer in their duration than those not combined with speech; and words uttered within a touch event were produced in a higher average pitch than those produced without touch [15, 16]. Finally, speech+touch events contained a potential signal for semantic matching in their multimodal temporal structure: when caregivers touched the part of the body they were speaking about, the speech and touch respective onsets and offsets were closely aligned, whereas no such alignment occurred when the caregiver was touching a body part they were not speaking about [15, 16].

4. Supervision and reinforcement versus prediction error

Infants' learning of body part words provides one powerful argument for some form of unsupervised learning where supervision and feedback are at most retrieved from prediction errors. Recall that in the [16] observational study, we asked parents to teach made-up words for body parts – for instance, for some caregiver-infant dyads, elbow was called "lepoga" and knee was called "dobita" (and the opposite for others). This design allowed us to make sure that infants had no prior exposure to these word forms. We observed that, when caregivers produced matching speech+touch input (i.e., saying *lepoga* while touching the elbow), infants looked more often at the target body part locations compared to other locations on their body, an effect that was absent when caregivers were producing mismatching speech+touch occurrences (i.e., saying *lepoga* while touching the knee). This learning was incredibly fast: Caregivers spent an average of 60 seconds on each word (range 15-180 seconds), and thus infants must have extracted at least a preliminary association, allowing them to distinguish matches and mismatches, in that timescale. As stated above, previous work in early language acquisition has demonstrated convincingly that labeled data (i.e., providing both semantic matches and mismatches labeled as such) and overt feedback (direct correction by adults) play a minor role in native human language acquisition. Our results of rapid learning in this study underline this point, as these infants seemed to have been learning these novel word-form body part associations likely unbeknownst to the adult interacting with them. Thus, if supervision and reinforcement play a role in learning, it is not on the basis of external categorizations,

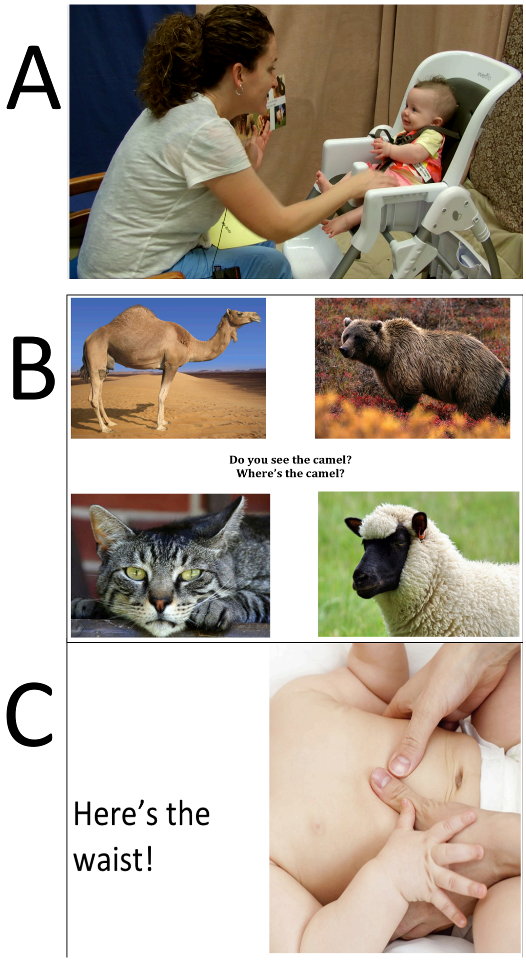

Figure 2: *Paradigm in Abu-Zhaya et al. (2016). Parents read 2 books, one with animal words, the other with body part words, in the setting portrayed in Panel A. Odd pages had a display of 4 pictures, the target word and 3 competitors; even pages had a blow-up of the target word. Panel B shows an example odd page from an animal book; panel C an even page from a body part book.*

but must instead be incorporated via some form of an internal mechanism, such as prediction error.

5. Future directions

5.1. Body part words could constrain syntactic acquisition

Throughout this paper, we pointed out aspects of infants' learning of body part words that are relevant to artificial multimodal learners. As we mentioned in the introduction, however, human learners not only learn a multimodal lexicon, but also embed it in a (proto-)syntax that constrains subsequent learning. For instance, by 14 months of age, infants learning French expect that novel words following *des* "the" and *ton* "your" will behave like nouns rather than verbs [17]. Since these clear indices for productive and flexible language models have been registered in infants much older than those discussed previously, we cannot be certain about whether there are any interesting effects for a system that starts out learning a proto-syntax from the distribution of body part words, names (the child's own name, "mommy"), and food words ("banana"), compared to later-emerging nouns that are more visually based ("cup"). Previous modeling work suggests that just a handful of items suffice to seed a process of discovery of syntactic lexical categories resulting in reasonable accuracy levels [18]. Nonetheless, it is possible that further work on protosyntax in younger infants that specifically targets the syntactic frames in which body part words occur may reveal additional insights that could be incorporated in artificial multimodal learners. Notice, for instance, that the syntactic distribution of names and common nouns is almost non-overlapping – thus already suggesting, from the very onset, a more complex categorization that infants and artificial learners could use to constrain their meaning discovery.

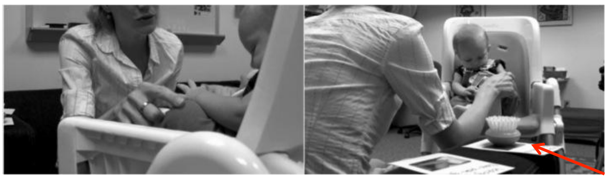

Figure 3: *Paradigm in Tincoff et al. (submitted). Caregivers taught the same syllable sequences used in the test phase of Seidl et al. (2015), associated with one of the following referents: brush (to which the red arrow points in the right panel), knee, and elbow. They were asked to teach these words "in whatever way feels most comfortable and fun for you and your baby." Their behavior and that of their child was captured through two cameras, as portrayed in the two panels.*

5.2. Multimodality and word-to-world mapping

One may wonder whether the arguments laid out above hold up when considering the other lexical types infants learn early in development. Results from laboratory studies examining American English-learning 6-month-olds' word-to-picture mapping show above-chance performance not only when infants are tested on hand versus feet [14], but also on mommy versus daddy [13], and a mixture of items largely drawn from food and body part words [12]. For each of these cases, how do infants segment the word form from the speech stream? How do they isolate the referent in the world to master the word-to-world mapping?

We have separately shown that American caregivers may provide aligned speech-touch temporal cues [15, 16], and that these well-timed touches aid infants' segmentation of syllable sequences when no other cues are available [8]. Although data on the other two types (food words and names) have not been similarly explored, we believe it is reasonable to hypothesize that these two lexical types may benefit from prosodic cues. We have argued previously that segmenting the word forms for proper names ("mommy" functions as a proper name in the infant world) is likely straightforward [8]: infants probably hear these names in isolation and/or in the vocative, in which case the word will be pronounced in the attention-grabbing tune. Similarly, in the culture where the infant experimental data were collected, mealtime is a time of active vocal exchange between caregivers and infants, and it is not unreasonable to imagine a parent saying "hmmm, banana!" effectively isolating this word in ways that may not be frequent in adult-directed interactions. In other work, we have seen that both proper names and food items are also segmented rather accurately by a range of word segmentation algorithms ran on transcripts of child-directed speech (see [19] for a description). These considerations could also pave the way for more effective artificial agents that attempt to build word form lexica based on unsegmented input, including agents that do not have access to many or all senses.

As for isolating the referent through the use of multimodal cues, we have discussed above how caregivers provide clear referential cues through touch in the case of body part words. Food words may similarly benefit from a rich confluence of cues, namely olfactory and gustatory ones. It remains to be seen whether other word types, which are part of the early lexicon, are placed in this lexicon (or not) depending on the richness of the individual's experience with the referent. For instance, perhaps "cow" is a later-acquired word for a city child, who has only seen this animal in books or on TV, but an early-acquired word for a child growing up on a farm, who has touched and smelled cows from early on. Most saliently, however, it is almost certain that body part words, proper names (mommy, the child's name), and food words are associated with some type of somatosensory experience, which could both enhance arousal and potentially aid in establishing the referent in the world. Thus, we believe it is still reasonable to posit that, in general, the early human lexicon is strongly grounded in the self. Hence, we would be interested in exploring whether artificial learners that profit from such principles would fare better than current models.

6. Conclusions

This paper summarizes experimental and observational work on very young human infants' word recognition and word learning. We have laid out a number of principles that guide and constrain this process, and purposefully drew from literature on infants aged 4 to 6 months to highlight the richness of the representations gained by these children, despite the very limited length of experience, and likely similarly limited processing power. We hope these results will inspire and challenge researchers building artificial learners.

7. Acknowledgements

AC gratefully acknowledges the Agence National de la Recherche (ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC).

8. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[2] A. Cristia and E. Dupoux, "Learnability differences in child- versus adult-directed speech: The case of unsupervised pattern discovery," *WIML, available from https://osf.io/qjxsu/*, 2016.

[3] J. Thomason, J. Sinapov, M. Svetlik, P. Stone, and R. J. Mooney, "Learning multi-modal grounded linguistic semantics by playing "i spy"," in *IJCAI*, 2016, pp. 3477–3483.

[4] A. Lazaridou, G. Chrupała, R. Fernández, and M. Baroni, "Multimodal semantic learning from child-directed input," in *Proceedings of NAACL-HLT*, 2016, pp. 387–392.

[5] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," *arXiv preprint arXiv:1206.6423*, 2012.

[6] C. Matuszek, L. Bo, L. Zettlemoyer, , and D. Fox, "Learning from unscripted deictic gesture and language for human-robot interactions," in *AAAI*, 2014, pp. 2556–2563.

[7] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.

[8] A. Seidl, R. Tincoff, C. Baker, and A. Cristia, "Why the body comes first: Effects of experimenter touch on infants' word finding," *Developmental Science*, vol. 18, no. 1, pp. 155–164, 2015.

[9] L. Gogate and G. Hollich, "Invariance detection within an interactive system: A perceptual gateway to language development," *Psychological Review*, vol. 117, no. 2, pp. 496–516, 2010.

[10] L. Smith and S. Gasser, "The development of embodied cognition: Six lessons from babies," *Artificial Life*, vol. 11, pp. 13–30, 2005.

[11] L. Smith, J. Maouene, and S. Hidaka, "The body and childrens word learning," in *The emerging spatial mind*, J. Plumert and J. Spencer, Eds. New York: Oxford University Press, 2007, pp. 168–192.

[12] E. Bergelson and D. Swingley, "At 6–9 months, human infants know the meanings of many common nouns," *Proceedings of the National Academy of Sciences*, vol. 109, no. 9, pp. 3253–3258, 2012.

[13] R. Tincoff and P. W. Jusczyk, "Some beginnings of word comprehension in 6-month-olds," *Psychological Science*, vol. 10, no. 2, pp. 172–175, 1999.

[14] ——, "Six-month-olds comprehend words that refer to parts of the body," *Infancy*, vol. 17, no. 4, pp. 432–444, 2012.

[15] R. Abu-Zhaya, A. Seidl, and A. Cristia, "Multimodal infant-directed communication: how caregivers combine tactile and linguistic cues," *Journal of Child Language*, pp. 1–29, 2016.

[16] R. Tincoff, A. Seidl, L. Buckley, C. Wojcik, and A. Cristia, "Feeling the way to words: Parents' speech and touch cues highlight word-to-world mappings of body parts," 2017, under review.

[17] R. Shi and A. Melançon, "Syntactic categorization in French-learning infants," *Infancy*, vol. 15, no. 5, pp. 517–533, 2010.

[18] A. Gutman, I. Dautriche, B. Crabbé, and A. Christophe, "Bootstrapping the syntactic bootstrapper: Probabilistic labeling of prosodic phrases," *Language Acquisition*, vol. 22, no. 3, pp. 285–309, 2015.

[19] E. Larsen, A. Cristia, and E. Dupoux, "Relating unsupervised word segmentation to reported vocabulary acquisition," *Interspeech*, 2017.