Language Acquisition is Embodied, Interactive, Emotive: a Research Proposal

Casey Kennington
Department of Computer Science
Boise State University
Boise, Idaho, U.S.A.
caseykennington@boisestate.edu

Abstract

Humans' experience of the world is profoundly multimodal from the beginning, so why do existing state-of-the-art language models only use text as a modality to learn and represent semantic meaning? In this paper we review the literature on the role of embodiment and emotion in the interactive setting of spoken dialogue as necessary prerequisites for language learning for human children, including how words in child vocabularies are largely concrete, then shift to become more abstract as the children get older. We sketch a model of semantics that leverages current transformer-based models and a word-level grounded model, then explain the robot-dialogue system that will make use of our semantic model, the setting for the system to learn language, and existing benchmarks for evaluation.

1 Introduction

Smith & Gasser (2005) showed that babies' experience of the world is profoundly multimodal: babies live in a physical world full of rich regularities that organize perception, action and thought. Babies explore the world in non goal-oriented ways, and babies learn in a social world to learn a shared linguistic communicative system that is symbolic. Indeed, a growing body of literature including child development, psychology, linguistics, and computational linguistics makes a strong case that the process of language learning (indeed, general human cognition) is embodied, interactive, and enacted (Pulvermüller, 1999; Lakoff & Johnson, 1999; Barsalou, 2008; Johnson, 2008; Smith & Samuelson, 2009; Di Paolo et al., 2018; Bisk et al., 2020). I argue that the *setting* (i.e., where and how the learning takes place) and stages of *progression* of how language is learned matters for holistic knowledge of semantic meaning, which has implications for how language is modeled computationally, especially in light of the fact that most language models, including recent transformer-based models like BERT (Devlin et al., 2018) and GPT-3 are derived abstractly from adult-written text.

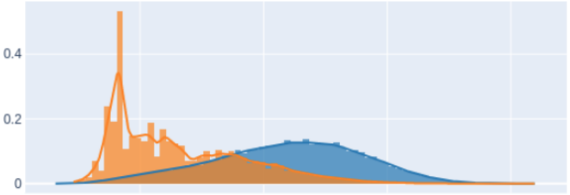

Figure 1: Average Age-of-Acquisition ratings for each entry in a subset of the WordNet dictionary: average ratings for word entries (blue) are higher than ratings for words in the definitions for those entries (orange).

Basic evidence that progression of learning matters is found in age-of-acquisition (AoA) datasets where known words are annotated with the average age when children are able to produce those words. Kuperman et al. (2012) presented ratings for over 30,000 English words (including nouns, verbs, and adjectives). For example the word *red*'s rating is 3.68 (i.e., 3 years + 0.68 towards the 4th year), and *abandon* is 8.32. Taking definitions from WordNet (Miller, 1995) for words that exist on the AoA dataset (totaling 26,919 words), the average AoA age for the words is 11 years (std 3.04), whereas the average AoA age for all of the words in the definitions is 6.59 (std 2.67). This is further illustrated in Figure 1, showing that words that make up definitions in the WordNet dictionary are learned much younger. This seems trivial; obviously words that are learned earlier in life are used to learn the meaning (or at least the definition) of words later in life, but language models and the data they leverage do not take this progression into account.

Vincent-Lamarre et al. (2016) gave this deeper consideration and found that recursively removing all words that are reachable by their definitions, but that do not define any further words, a dictionary can be reduced to 10% of its size. This can further be reduced to a strongly connected subset of words that all other words can be defined from, which comprises only about 1% of the dictionary. This is important because while many words are defined by other words, there is a core subset of words that cannot be defined by other words (e.g., what is the dictionary definition of the word *red*?), but rather must be experienced directly, otherwise meaning of all words are completely ungrounded; "just strings of meaningless symbols (defining words) pointing to meaningless symbols (defined words)" which is precisely what the symbol grounding problem is (Harnad, 1990).[1] Vincent-Lamarre et al. (2016) showed that the "core" words, from which all other words are eventually defined, are learned earlier in life and they are more *concrete*; i.e., they are words that denote physical, tangible objects and proprioperceptive embodied states. This isn't to say that humans don't have the capacity for abstraction without concrete experience. Indeed, all words are to some degree abstract because they make an abstract categorization, even ones that directly denote perceptual experience (Harnad, 2017). But given the literature cited above, no human would likely arrive at abstract thought without something to abstract over, i.e., words that denote concrete, physical entities.

Requirements Despite important advances for natural language processing (NLP) tasks and applications, it is clear that models trained only on text are missing critical semantic information (see, for example, Rogers et al. (2020) which reviews relevant literature). These kinds of text-only models make an *abstractness assumption* because the only "context" they make use of is lexical context (i.e., words are only "defined" by other words), whereas a holistic model of meaning requires lexical, embodied (including emotion, see Lane & Nadel (2002); Moro et al. (2020)), perceptual (i.e., connected to the world–symbol grounding), and interactive (i.e., conversational grounding (Clark, 1996)) context in the language learning process.

2 Research Proposal

The goal of this research project is to work towards a model of semantic meaning that handles concrete and abstract meaning acquisition, encodes emotion, and is learned in a setting similar to that of a child: embodied, spoken interaction.

Embodiment The semantic model needs to be housed in a physical body that can perceive and act in the world. We opt to use robotic platforms, beginning with Anki Cozmo which has been shown to be suitable for the setting of first-language acquisition because people perceive Cozmo to have a young age (Plane et al., 2018), though clearly Cozmo is nothing like a real child in its sensory capabilities or affordances. Cozmo's perceptual abilities include camera input, we add an external microphone, and tracking of internal state variables (e.g., lift height, head angle, wheel speed) and Cozmo's abilities for action include driving forward and backward, turning, lifting and lowering a small lift arm that can move specific types of objects, as well as up and down movement of the head and speech synthesis with a young-sounding voice.

Interaction The setting for the robot to learn language is face-to-face spoken, interactive dialogue which is the basic setting where humans learn their first language (Fillmore, 1981). This *situated* setting requires a spoken dialogue system (SDS) that is well-suited for robots. Following Kennington et al. (2020), such a robot-ready SDS must be (1) *modular* so it can integrate with various robot modules, (2) *multimodal* the semantic model should incorporate perception (including proprioperception), (3) *distributive* so robot modules and SDS modules can communicate with each other across distributed hardware, (4) *incremental* so processing can happen quickly and immediately, and (5) *aligned* in that sensors and actions must be synchronized temporally. The incremental requirement is crucial: the semantic model must not wait for full, grammatical utterances; rather, it

[1] Note that the author of Harnad (1990) is an author of Vincent-Lamarre et al. (2016), which makes these claims.

should process by word (or sub-word) increments because humans process spoken input in real-time (Eberhard et al., 1995), though many models of semantics require full, sentence-level input.

Emotion The meanings of many words have emotion as part of their connotation (Lane & Nadel, 2002) and that emotion plays a role in the meaning of abstract words (Vigliocco et al., 2014). Following Moro et al. (2020), similar to semantic word meaning, emotions can be viewed as on a continuum between abstract and concrete; abstract according their lexical categories (e.g., *happiness*, *fear*, *anger*) distributed with text (Barrett, 2017), and concretely through *affect* which is a biological system and a fundamental part of embodiment (Vigliocco et al., 2014). In contrast to abstract emotion concepts, affect is a more basic underpinning for emotion, ranging from unpleasant to pleasant (valence) and from agitated to calm (arousal), which, like vision, is something that could potentaily be grounded into if a model exists. We use the model we introduced in (McNeill & Kennington, 2019) as a proxy for affect as it maps from robot behaviors to a distribution over 16 affects that were labeled by humans who observed the behaviors for affective display.

Open Questions This research highlights some important questions relating to semantic meaning of language, how it is learned, the role of emotion in meaning (and language acquisition) which has implications for NLP, robotics, human-robot interaction, and artificial intelligence applications. Specifically, we ask the following questions:

- What semantic model fulfills the requirements of being grounded, can learn meanings of words with only a few examples (i.e., fast-mapping as children can do) or directly from explanation, and can be learned through interaction with others and with the physical environment?
- How can concrete and abstract meaning be learned and represented in that model?
- Can affect and emotion help reconcile the concrete and abstract meaning learning and representations?

Model Sketch Two models that inspire our proposed model (though there are many other vision-lanuage models) are (1) VilBERT (Lu et al., 2019), a dual-transfomer architecture that brings together textual embeddings and images for image description generation, more recently leveraged for visual dialogue (Murahari et al., 2019), and (2) the *words-as-classifiers* (WAC) model (Kennington & Schlangen, 2015), a simple grounded modal that amounts to a binary classifier for each word, each classifier yields a “fitness” score, given a representation of a visual object and word in question. Schlangen et al. (2016) use the WAC model with images of real objects in a reference resolution task, using vectorized representations of objects from a convolutional neural network trained on imagenet data. WAC has been used as a grounded model for modalities beyond vision; for example, Moro et al. (2020) grounded WAC into low-level affect predictions which included robot internal states and audio representations.

Both models have their advantages. VilBERT is transformer-based and leverages BERT for representing language, making it robust to various language input. Being a word-level model, WAC has the advantage of being useable in an incremental SDS setting–a requirement for robot-ready SDS–and can learn fitness scores with only a few training examples. For example, McNeill & Kennington (2020) recently used WAC in an interactive language acquisition study using WAC as the semantic model on the Cozmo robot with human participants; WAC was able to quickly learn a handful of vocabulary words despite the short interactions. Both models have disadvantages. VilBERT, like most neural models, is data hungry and uses BERT which is trained on text, yet children learn interactively and often with only a small set of examples. Moreover, VilBERT is currently strictly designed to model the visual modality taken from images. WAC suffers from two strong assumptions. First, that all words are trained and applied independently from each other (making WAC’s composition strategy quite limiting–it simply multiplies the fitness scores together to form sentence-evel scores for referring to objects) and second, that all words fully denote concrete things, despite many words being abstract and therefore do not have physical manifestations (e.g., *utopia* or *beneficial*).

Our current work explores using WAC classifiers as embeddings for the VilBERT model by training WAC classifiers (i.e., simple multi-layer perceptrons) on images using positive and negative image examples for each word, then extracting the coefficients of those classifiers (i.e., concatenate the coefficients for each layer, forming a vector) that we then use as input embeddings for the language

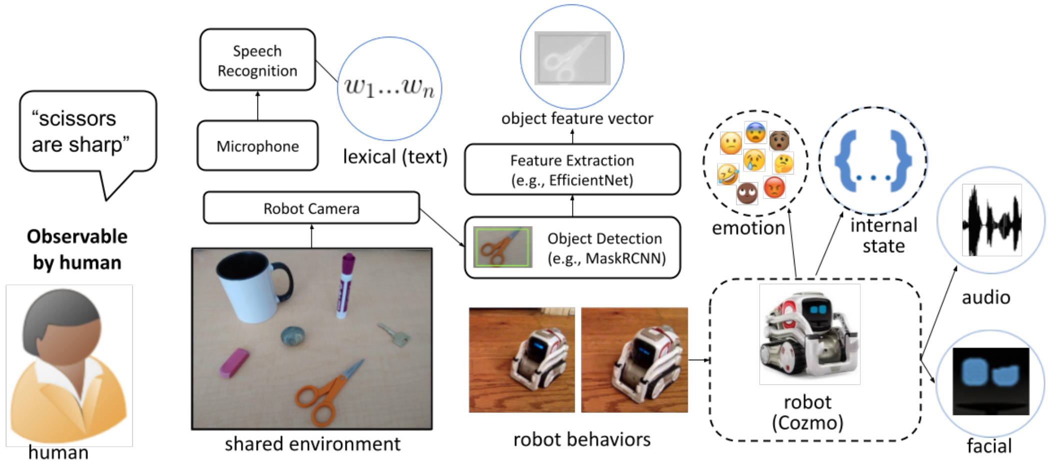

Figure 2: Overview of proposed system. The model takes in 6 modalities (depicted in the circles); modalities come from the user's speech, shared environment, or robot. Emotion is grounded into by the model, and (including emotional) robot behaviors facilitate the interaction. The user utterance contains concrete and abstract terms.

side of the VilBERT model. Current results show promise on the visual dialogue benchmark (Murahari et al., 2019) that this is a useful addition to the VilBERT model, thereby unifying the two models and overcoming some of the assumptions and shortcomings of each model used alone, but more work is needed for a model that fulfills all of the requirements and works for a robotic platform in a spoken dialogue setting.

System We will use the system for incremental dialogue described in Kennington et al. (2020), which has modules for speech recognition, object detection using YOLOv4 (Bochkovskiy et al., 2020), and object feature representation by using the topdropout layer from EfficientNet (Tan & Le, 2019) to feed into our semantic model. The rrSDS platform has bindings for the Cozmo robot, as well as bindings for OpenDial (Lison & Kennington, 2016), which we will use for dialogue act and robot action decisions. We will use the model described in McNeill & Kennington (2019) to map from robot modalities to a distribution over a representation of affect that our model of semantics will ground into. Our system is portrayed visually in Figure 2, including the Cozmo robot.

Evaluation Plan We will recruit human participants and task them with interacting with Cozmo first by referring to concrete objects (e.g., *scissors near you*, as done in McNeill & Kennington (2020), then task them with engaging in more abstract comparisons or truth claims (e.g., *scissors are sharp*). To gain exposure to larger vocabulary, we will put Cozmo in varied contexts with participants where participants also interact with Cozmo at intervals across a long time span. We hypothesize that this will result in a representation of semantic meaning encoded in a model like VilBERT that has higher fidelity to the setting and circumstances in which human children learn language. We will continue our ongoing work in using this model on known benchmarks (e.g., GLUE (Wang et al., 2018), visual dialogue (Murahari et al., 2019)) to compare with existing models that are only trained on text data.

3 Conclusions

The process whereby human children learn language is vastly different from the process whereby existing state-of-the-art language models learn language. While current advancements in multimodal language grounding are moving in the right direction, the setting (i.e., situated dialogue) and lack of embodiment still pose a challenge. Addressing these requirements in a single system is by no means low-hanging research fruit; it requires interdisciplinary background in NLP, SDS, and human-robot interaction research, but we believe that the efforts will be beneficial in working towards natural communication with automated systems, including robots.

References

Lisa Feldman Barrett. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.

Lawrence W Barsalou. Grounded Cognition. *Annual Review of Psychology*, (59):617–645, 2008. doi: 10.1146/annurev.psych.59.103006.093639. URL http://psych.annualreviews.org.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience Grounds Language. *arXiv*, 2020. URL http://arxiv.org/abs/2004.10151.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.

Herbert H Clark. *Using Language*. Cambridge University Press, 1996. ISBN 0521567459. doi: 10.2277/0521561582. URL http://www.amazon.com/Using-Language-Herbert-H-Clark/dp/0521567459.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. oct 2018. URL http://arxiv.org/abs/1810.04805.

Ezequiel A Di Paolo, Elena Clare Cuffari, and Hanne De Jaegher. *Linguistic bodies: The continuity between life and language*. Mit Press, 2018.

Kathleen M. Eberhard, Michael J. Spivey-Knowlton, Julie C. Sedivy, and Michael K. Tanenhaus. Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24(6):409–436, 1995. ISSN 00906905. doi: 10.1007/BF02143160.

Charles J. Fillmore. Pragmatics and the description of discourse. *Radical pragmatics*, pp. 143–166, 1981.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990. ISSN 01672789. doi: 10.1016/0167-2789(90)90087-6. URL http://eprints.soton.ac.uk/258175/1/sgproblem1.html.

Stevan Harnad. To cognize is to categorize: Cognition is categorization. In *Handbook of Categorization in Cognitive Science*, pp. 21–54. 2017. ISBN 9780081011072. doi: 10.1016/B978-0-08-101107-2.00002-6. URL http://www.unites.uqam.ca/sccog/liens/program.html.

Mark Johnson. *The meaning of the body: Aesthetics of human understanding*. University of Chicago Press, 2008.

Casey Kennington and David Schlangen. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 292–301, Beijing, China, 2015. Association for Computational Linguistics. ISBN 9781941643723. URL http://www.aclweb.org/anthology/P15-1029.

Casey Kennington, Daniele Moro, Lucas Marchand, Jake Carns, and David McNeill. rrSDS: Towards a Robot-ready Spoken Dialogue System. In *Proceedings of the 21st Annual SIGdial Meeting on Discourse and Dialogue*, Virtual, 2020. Association for Computational Linguistics.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990, 2012. ISSN 15543528. doi: 10.3758/s13428-012-0210-4. URL https://www.mturk.com/mturk/welcome.

George Lakoff and Mark Johnson. *Philosophy in the flesh: The embodied mind and its challenge to western thought*, volume 640. Basic books New York, 1999.

Richard D Lane and Lynn Nadel. *Cognitive Neuroscience of Emotion*. Oxford University Press, 2002.

Pierre Lison and Casey Kennington. OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - System Demonstrations*, 2016. ISBN 9781510827615.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. 2019. URL `https://arxiv.org/pdf/1908.02265.pdfhttp://arxiv.org/abs/1908.02265`.

David McNeill and Casey Kennington. Predicting Human Interpretations of Affect and Valence in a Social Robot. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany, jun 2019. doi: 10.15607/RSS.2019.XV.041.

David McNeill and Casey Kennington. Learning Word Groundings from Humans Facilitated by Robot Emotional Displays. In *Proceedings of the 21st Annual SIGdial Meeting on Discourse and Dialogue*, Virtual, 2020. Association for Computational Linguistics.

George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 1995. ISSN 15577317. doi: 10.1145/219717.219748.

Daniele Moro, Gerardo Caracas, David McNeill, and Casey Kennington. Semantics with Feeling: Emotions for Abstract Embedding, Affect for Concrete Grounding. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Virtual, 2020.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *arXiv preprint arXiv:1912.02379*, 2019.

Sarah Plane, Ariel Marvasti, Tyler Egan, and Casey Kennington. Predicting Perceived Age: Both Language Ability and Appearance are Important. In *Proceedings of SigDial*, 2018.

Friedemann Pulvermüller. Words in the brain's language, 1999. ISSN 0140525X.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What we know about how BERT works. *arXiv*, 2020. URL `https://github.com/`.

David Schlangen, Sina Zarriess, and Casey Kennington. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1213–1223, 2016. ISBN 9781510827585. URL `http://arxiv.org/abs/1510.02125`.

L B Smith and L Samuelson. Objects in Space and Mind: From Reaching to Words. In *The Spatial Foundations of Language and Cognition*, 2009.

Linda Smith and Michael Gasser. The Development of Embodied Cognition: Six Lessons from Babies. *Artificial Life*, (11):13–29, 2005.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv*, may 2019. URL `http://arxiv.org/abs/1905.11946`.

Gabriella Vigliocco, Stavroula Thaleia Kousta, Pasquale Anthony Della Rosa, David P. Vinson, Marco Tettamanti, Joseph T. Devlin, and Stefano F. Cappa. The neural representation of abstract words: The role of emotion. *Cerebral Cortex*, 2014. ISSN 14602199. doi: 10.1093/cercor/bht025.

Philippe Vincent-Lamarre, Alexandre Blondin Massé, Marcos Lopes, Mélanie Lord, Odile Marcotte, and Stevan Harnad. The Latent Structure of Dictionaries. *Topics in Cognitive Science*, 8(3):625–659, jul 2016. ISSN 17568765. doi: 10.1111/tops.12211. URL `http://doi.wiley.com/10.1111/tops.12211`.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018. doi: 10.18653/v1/w18-5446.