The Symbol Grounding Problem Re-framed as Concreteness-Abstractness Learned through Spoken Interaction

Casey Kennington
Boise State University
1910 W University DR
Boise, ID 83725
caseykennington@
boisestate.edu

Osama Natouf
Boise State University
1910 W University DR
Boise, ID 83725
osamanatouf@
u.boisestate.edu

Abstract

The Symbol Grounding Problem points out that the underlying mechanisms of computation are symbolic and, therefore, missing crucial information when they are used for processing natural language until they are somehow able to perceive the world directly. Our goal in this paper is twofold: First, we review some of the recent literature that claims to address (even if just to a small degree) the Symbol Grounding Problem, and explain why it is still yet a problem partially due to a misinterpretation of the problem and that there are more modalilties that symbols need to ground into beyond just pictures, including emotion. Second, we re-frame the problem as a problem of handling concreteness and abstractness because (perhaps surprisingly) computational models of distributional meaning seem to capture abstractness more directly than they do concreteness. We take inspiration from child development and offer a toy example of how one could approach modeling concrete and progressively more abstract words. We conclude by posing some open questions and offering paths for future work.

1 Introduction

The *Symbol Grounding Problem* posits that linguistic meaning cannot be directly encoded in a computational symbol, particularly because the meanings of many words are *grounded* in real-world experience (Harnad, 1990). For example, the word *blue* is a color, but so is *red*, so knowing that they are in the same category of words does not uncover their meaning because both denote different swathes of the color spectrum that is visible to humans, and without experiencing each word used in physical contexts of other people denoting those colors, it is impossible to learn each word's connotation.

Harnad (1990) identified properties of symbolic systems; for example that there are atomic symbols and composed symbol combinations, and that symbols (can be) semantically interpretable, but this is in contrast to how humans can discriminate, manipulate, identify, and describe objects, and humans can even respond to the objects and descriptions of those objects. Put another way, humans interact with and talk about the world, and the cognitive capabilities that humans have are a result of the fact that they do so (Smith and Gasser, 2005).

A recent neurological study gives empirical backing to this proposition where the authors "assessed the extent to which different representational systems contribute to the instantiation of lexical concepts in high-level, heteromodal cortical areas previously associated with semantic cognition" (Fernandino et al., 2022). Their work showed that, though semantic information can be represented by distributional representations and symbolic taxonomies, a clear advantage exists for "experiential representational structures" such as sensory-motor, affective, and other features of phenomenal experience, suggesting that if research is to solve the problem of acquiring, representing, and applying linguistic meaning computationally, then to learn a word's semantics means access to experience with the world.

More than 30 years have elapsed since Harnad (1990), now with over 5,000 citations which suggests, at the very least, that the problem has been considered and taken seriously by scientists. However, fifteen years after the original publication, Taddeo and Floridi (2005) reviewed the literature of the time and concluded that at that point, the problem was far from solved. Since then, the "collectivist" models that seemed promising at the time have evolved to deep learning models that have proven their power on language tasks, with some models showing promise on language and vision tasks. Does this mean the problem has been solved, or will be soon, given the right deep learner?

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue,
August, 22-24, 2022, Dublin.

In this paper, we explore some of the recent work on symbol grounding. We observe that the way language is currently modeled suffers from a similar problem that symbolic systems suffered: they are ungrounded (Section 2). Moreover, though vision is an important modality for symbol grounding, it is not the only important one (including emotion). In Section 5 we identify other modalities that are often ignored, but must be part of any model that claims to be holistic. However, not all words need to be grounded into in order to arrive at their meaning; we therefore re-frame the Symbol Grounding Problem in light of an important distinction between concrete and abstract concepts in Section 3 which, we believe, have implications for how meaning can be modeled in existing deep learners. We conclude by offering some suggestions for avenues of future research.

2 The challenge of symbol grounding

Harnad pointed to Searle (1980)'s Chinese Room as a metaphor which challenges the core assumptions that symbols carry meaning on their own. He explains that if he, someone who could not read or speak Chinese, were in a room with a Chinese-Chinese dictionary and had the instuctions to take "input" of one Chinese character, look up the character in the dictionary and then find the "output" character, even if the inputs and outputs were perfectly mapped as observed by an outsider, the person in the room doesn't actually know Chinese, which is the same problem that computers have when they process natural human language.

Yet are words not also symbols? In some ways yes, but we need to be clear here what is meant by *word*. A word is a linguistic unit that carries linguistic meaning on its own and can be used as a placeholder for a concept much like symbols can. For example the word *chair* can denote real chairs, but uttering or writing the word can replace the presence of chairs when someone wishes to talk about the concept of a chair—the word *chair* effectively becomes an abstraction of the connotation. The confusion comes when one assumes that the word *chair* as it is written actually represents the concept itself, but it does not; the concept of *chair* resides in human brains, but because written text is computable and since text is a placeholder for concepts for humans as they communicate with each other, it follows that machines could use text as symbols and text would carry the meaning, but that is precisely what the Symbol Grounding Problem is pointing out does not work because, like symbols, text is ungrounded.

Since 1990, other models of learning and representing linguistic meaning that go beyond the kinds of symbols that Harnad was referring to, most notable embeddings and language models that follow the *distributional hypothesis*, a hypothesis that posits that the meaning of a word can be derived by how it is used in the context of other words within text; the Firthian "you shall know a word by the company it keeps" generally means that words keep company with other words. This led to models such as word embedding vectors (Mikolov et al., 2015) and, more recently, powerful transformer-based language models like BERT (Devlin et al., 2018) that are trained on text alone; the training regime is often a task of guess-the-masked-word in a context of other words. These models both in their time have revolutionized entire research fields. Have they solved the Symbol Grounding Problem?

With attention now to language models instead of symbols, and building on Searle's Chinese Room thought experiment, Bender and Koller (2020) argue that language models do not learn meaning on similar grounds as the Symbol Grounding Problem. They offer *the octopus test* where an octopus "overhears" a conversation between two people on desert islands by tapping into the communication wire that connects them. The octopus learns how to mimic one of the dialogue partners by learning regularities in the kinds of words and phrases they use, and when the octopus has an opportunity to take over the role of that particular dialogue partner, despite being able to learn patterns of words and how they should appear in the context of other words, the octopus cannot answer simple questions because the octopus fails to know the kinds of objects that certain words denote. In other words, despite their success, models that follow from the distributional hypothesis also fail at solving the Symbol Grounding Problem. Furthermore, Herbelot (2013) makes a strong case that text alone cannot possibly be expected to contain the meaning of many words, no matter how much text is used for training. Clearly, however, some degree of meaning can be derived and represented from text, otherwise language models could not possibly work so well on so many natural language processing tasks, which begs what kind of information they are learning (see Rogers et al. (2020) for a review), though it is clearly not grounded.

Dictionaries, likewise, do not solve the Symbol Grounding Problem even if each word in the dictionary has a corresponding definition that is intended to represent meaning of words, or at least the description of the meaning of words. Harnad explicitly mentioned the "dictionary merry-go-round" of words defining other words, a claim that was empirically tested in Vincent-Lamarre et al. (2016) (work by Harnad and colleagues) who identified a subset of words that all other words are eventually defined by, showing that defining words by other words is indeed useful, but do not capture holistic meaning. Conversely, not all words need to be grounded—meaning can be derived from other words in many cases. The challenge is determining which word meanings that should ground into the physical world and which word meanings that should be derived from lexical context (i.e., text).

3 Reframing the problem: concreteness & abstractness

We argue that framing the Symbol Grounding Problem as a question of *concreteness* vs. *abstractness* puts the research field on better theoretical footing to make the best of what is required for solving the Symbol Grounding Problem and existing computational models that derive meaning from distributional approaches using text. In this section, we explain and give examples of concreteness and abstractness, argue that no current model captures both, and perform a small scale toy experiment to explore what a model that does capture both might look like.

Concrete words are words that denote physical things like objects, shape, and color (e.g., *chair*, *red*), requiring Symbol Grounding to arrive at meaning, whereas abstract words are words that denote ideas (e.g., *democracy*, *travel*), but it should be noted that the distinction between concrete and abstract concepts lies on a continuum, not a binary dichotomy (Della Rosa et al., 2010; Brysbaert et al., 2014). Thus some words are more concrete or abstract than others, some examples that illustrate this are shown in Figure 1. Words range from very concrete (e.g., *ball*) to very abstract (e.g., *utopia*). For more concrete words, corresponding images show clear examples of something that the word can denote visually. However, more abstract words can have aspects of their meaning represented visually, but not fully (e.g., *democracy* includes voting, but voting is only one aspect of the meaning of *democracy*).

That some words need grounding while others do not begs the question *Which words need symbol grounding?* Words that are more concrete like *ball* and *red* clearly need to be grounded. The word *red*, for example, can be understood to some degree without grounding, for example that it is a color and that certain objects can be red (e.g., apples and vehicles), and while it is true that there are metaphorical uses for the word *red*, those metaphorical uses can only be understood after knowledge about *red* as a color is learned (see arguments made in Lakoff and Johnson (2008) about metaphors; see also Bizzoni and Dobnik for discussion on visually grounded metaphors). A fairly simple grounding strategy could be used at the word level to arrive at a grounded representation, for example Schlangen et al. (2016) where each word in a corpus was represented by a binary classifier; the inputs of which were visual features. The model, however, assumes that all words are in fact concrete and visually grounded.

On the other end of the continuum are abstract words like *democracy* and *utopia*. Even though someone could imagine a visual depiction of either of those terms, their meaning is not grounded directly into the physical world, but are rather ideas that are defined by other words. Because the meaning of abstract words can be defined by other words, it is the meaning of abstract words that is captured by distributional methods, such as recent language models like BERT (Devlin et al., 2018). Distributional approaches, as noted above, are trained on text and make the tacit assumption that all words are abstract—ungroudned—even words that show up in the text that are in reality concrete are assumed model to be abstract in how the model captures meaning.

Is there a model that can capture both concreteness and abstractness? The real challenge comes from words that are not obviously concrete nor obviously abstract, rather somewhere in between like *farm*, or *color*. A farm can be observed and denoted visually, but what makes a farm a farm is not represented by an image or a series of images, but rather specific (abstract) properties like growing crops or keeping livestock within a specified land area. However, the words that are required for one to understand the concept of *farm*, one must understand what crops, land, and livestock are, concepts that

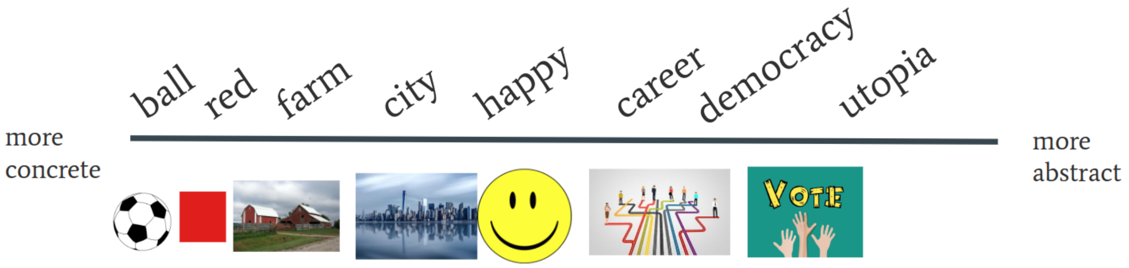

Figure 1: Examples of words that are more concrete vs. more abstract. Words that are concrete have physical (in this case, visual) denotations, whereas more abstract words do not physically exist. Concreteness ratings from Brysbaert et al. (2014) resulted in the placement of the words.

are themselves to some degree more concrete and grounded. The other example, *color* might be more illustrative: as a concept, *color* seems concrete because it is a very visual concept that categorizes colors.[1] The meaning of the word *color* can be defined by other terms, but the function of the word itself is to distinguish between other words that are considered colors like *red* and *blue* and words that are not. Thus while *color* itself does not directly ground into the visual world, it does directly connect somehow to words that in turn are grounded in the visual world. So should *color* be learned as an abstract concept or one that categorizes concrete concepts? In the following section we explore the latter with a simple toy example using a handful of categories and related words for each category.

3.1 A toy experiment: grounding into concrete words meanings

We conduct here a small experiment to test the possibility that concrete words can be "grounded into" by more abstract words that are higher on the abstractness scale, where what is grounded into differs depending on the level of abstraction. We use the following train and test set vocabularies for five "abstract" categories (i.e., not fully concrete); each item begins with with a more abstract word in boldface that is a grouping of the other words, which are all more concrete. Note that none of the test words are also in the training set.

Train:

- **color**: red, blue, green, yellow, brown
- **animal**: dog, cow, cat, mouse, bird
- **furniture**: couch, chair, desk, bed
- **vehicle**: car, van, truck, pickup, tractor
- **appliance**: stove, oven, microwave

Test:

- **color**: orange, purple
- **horse**, sheep
- **furniture**: table, sofa
- **vehicle**: taxi, jeep
- **appliance**: mixer

Procedure Following the words-as-classifiers (WAC) approach to grounded semantics (Schlangen et al., 2016), we train a logistic regression classifier (C=0.25, max-iter=1000) for each concrete word using images (we downloaded top 100 images for each word as a search term using Google Image Search) that have been passed as input into the CLIP model (Jia et al., 2021) which yields a vector of size 512 for each image. Negative examples of each word are randomly sampled from images for other words; we use three negative examples for each positive example. This results in a trained binary classifier for each concrete word that can, given a new image (i.e., represented as a CLIP vector), determine how well a trained classifier for a word *fits* the image. This is depicted in the top portion of Figure 2. For example, a trained classifier for *red*, given an image with a lot of red in it, would return a higher probability than if the image had little or no red in it. With our toy example, we therefore have 30 trained classifiers for each of the concrete words in both the train and test sets.

We then train classifiers for the more abstract words (i.e., *color, animal, furniture, vehicle, appliance*) that are defined by how they group together corresponding concrete words in a similar

[1] The concreteness rating for *color* in Brysbaert et al. (2014) is 4.08, which makes it a fairly abstract, compared to *democracy* (1.78) and *chair* (4.58)—higher numbers denote higher concreteness. *Color* is only slightly more abstract than *chair*.

way, though using different features. We hypothesize that the groupings are based on the feature sets that are common to the different categories. We therefore use the *coefficients* of the trained concrete classifiers as input to the abstract word classifiers because, as pointed out by Schlangen et al. (2016), the trained classifiers themselves (which are the logistic function and corresponding coefficients) represent a computational intension of each word; positive examples are the words listed for a category, negative examples randomly sampled from from the other categories; three negative examples for each positive. This is partially depicted in Figure 2 that shows how coefficients from the trained *red* classifier are features for the *color* classifier.

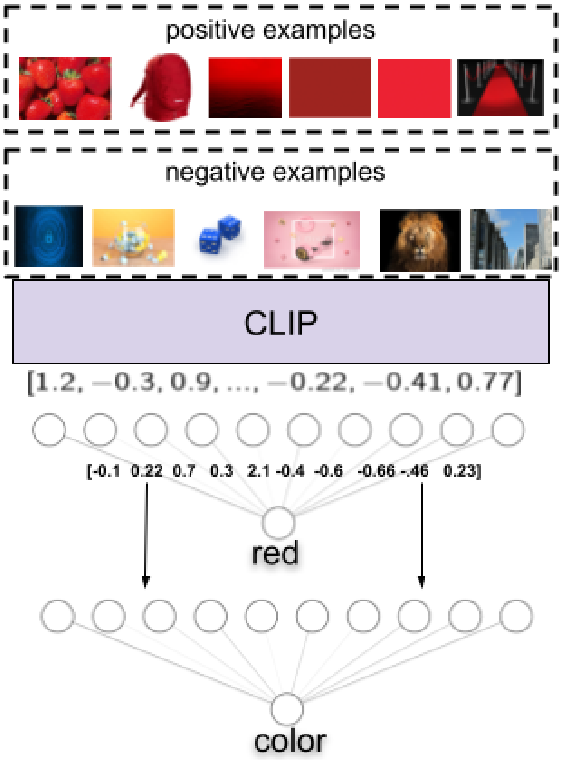

Figure 2: Example of *red* WAC classifier trained on positive negative examples of images represented by CLIP vectors. The coefficients for the *red* classifier are a positive input for the *color* classifier. Figure adapted from Kennington (2021) with permission from the author.

Task & Results We task the five trained abstract category classifiers to correctly classify the words in the test set, resulting in an accuracy metric. To evaluate, we take the concrete words for the test set, then pass their coefficients as input to each of the five classifiers trained for each category and compare the category with the highest probability to the labeled category. The test set only has nine words, but even with a small training set, our evaluation yields 88% accuracy (the only mis-classified word was *mixer* which was mis-classified as *furniture* instead of *appliance*).

This toy experiment shows that it could be possible to build meanings of words that are somewhat abstract by grounding them into words that are concrete by treating coefficients as a level of abstraction, which may solve the grounding problem for some of the words that are closer to the concrete side of the concrete-abstract continuum. One major limitation is that the model assumes that all words are independent of each other in how they are trained; e.g., the word *color* has no knowledge about *appliance* even though appliances can have color—knowledge that could be picked up from text using distributional methods. Another limitation is determining which words are used as positive examples to a more abstract word. In the toy dataset above, the categories are clear, but it is unlikely that an abstract word's meaning can be derived from the coefficients of the words in its dictionary definition as positive examples. In the following section, we explore how one might arrive at a model that does learn from concrete to abstract over time by taking inspiration from child development.

4 Learning meaning from concrete to abstract

With the Symbol Grounding Problem recast as a problem of grounding for concrete words directly into perception, and more abstract words into some kind of meaning representation of more concrete words, and building on the successes of the distributional approaches to modeling language, we now consider what it takes to learn concrete words, followed by more abstract words that build upon them. To do so, we take inspiration from child development where the setting of spoken interaction is crucial to learning concrete terms initially, and how emotion is integral to the process of learning language and is part linguistic meaning itself.

4.1 The setting of spoken interaction

Complementary to Symbol Grounding, *Communicative Grounding* is the process of mediating what words mean between an individual and a language community through active use of language (Clark, 1996). Communicative Grounding is crucial to Symbol Grounding because, as argued by Larsson (2018), Symbol Grounding is a side effect of Communicative Grounding. To illustrate: if two individuals are sitting together in a park and actively observe a kite in the air, and one person utters *kite*, the other person who had never seen a kite before now grounds the word *kite* with the observed object (symbol grounding), and both individuals know that each other has taken part in the interaction (communicative grounding).

Following Kennington (2021), children learn their first language in this highly interactive setting where communicative grounding between caregiver and child takes place as a facilitator for Symbol Grounding, and speech is the primary modality of linguistic interaction. At this early language learning stage, children generally learn words that denote physical objects making them largely concrete (Kuperman et al., 2012; Clark, 2013; Borghi et al., 2019; Ponari et al., 2018). Furthermore, Locke (1995) makes a case that that putting an agent (or, we conjecture, a computational model) in a place where it can only *observe* language—be it text or even referring expressions made to visually present objects—does not bring the child (or a computational models) to language capabilities as much as *participatory interaction*.[2] Before children can comprehend or utter words that carry semantic content in a given language, they experience the world in a profoundly multimodal and interactive setting (Smith and Gasser, 2005), giving children existing experience with the physical world that they can later leverage when learning their first words by categorizing perceptual input and grounding word concepts to those groupings.

Moreover, children tend to move in a learning progression of concrete to abstract over time: Borghi et al. (2019) notes that data indicate that only 10% of the vocabulary of 4-year-olds is composed of abstract words, abstract words represent 25% of 5-year-olds' words and more than 40% of 12-year-olds' vocabulary (see also Ponari et al. (2018)). Put succinctly, the words that children first learn largely require symbol grounding, but meanings of later words that are more abstract can be learned by how they are defined by and used with other words distributionally. This is not to claim in any way that children only learn concrete words early in life, then move to learning only abstract words—humans learn new concrete terms throughout life, and children begin to learn fairly abstract concepts early in development (e.g., greetings). Furthermore, this is not to say that cognition is purely a bottom-up process; clearly there is some degree of cognitive processing that is top down—the natural process of categorization of sensory input is an integral part of cognition whether the categories are innate or not (Harnad, 2017).[3]

4.2 Concrete-affect; abstract-emotion

Missing from the discussion thus far in language learning—both concrete and abstract—is how emotion plays a role that works in parallel to the concrete-to-abstract language learning progression. Early on in a longitudinal project (Alan Sroufe et al., 2009), the authors note that cognitive advances "promote exploration, social development, and the differentiation of affect; and affective-social growth leads cognitive development [...] neither the cognitive nor the affective system can be considered dominant or more basic than the other; they are inseparable manifestations of the same integrated process [...] It is as valid to say that cognition is in the service of affect as to say that affect reflects cognitive processes." In other words, cognition is not disconnected from emotion. Locke (1995) agrees, while tying emotion directly to language: in the real speech of sophisticated speakers, where both linguistic content and vocal affect are present, one type of cue does not preempt the other–and for speech to work this must be the case. Listeners must know both what the speaker is saying and what they intend by saying it. Humans duplexly pick up information about the linguistic content *and* the speaker's affect because the cues to these things are of different sorts and are processed by different brain mechanisms—this is particularly important for children who are learning their first words. Thus, according to Locke, the meaning of an utterance is in the linguistic content, but the *intent* of the speaker who made the utterance is also in the affect and emotion. In fact, children are adept at reading intents of others via affect and emotion, before they can even speak or really understand words (Smith and Gasser, 2005). This

[2] Sachs et al. (1981) explained that two children with normal hearing were born to deaf parents, so the parents did not use speech interaction with their children. Despite watching television with programming for children, their speaking abilities were far behind their peers, which required intervention.

[3] Missing from this discussion is how *affordances* affect perception and categorization, but note that understanding object affordances are an important part of the concept learning process.

suggests that emotional states exist within humans before they can speak; indeed, emotions can facilitate the language learning process for someone who is learning their earliest words (McNeill and Kennington, 2020).

Furthermore, recent empirical work in neuroscience and cognitive science have explored the relationship between language and emotion. Lane and Nadel (2002) explained that the meaning of many words has emotion as part of their connotation, and Mazzuca et al. (2018); Villani et al. (2021) have shown that abstract linguistic concepts are more closely tied to emotion (i.e., interoception) in particular emotional and mental states, as well as social concepts than concrete linguistic concepts are tied to emotion. Moreover, Ponari et al. (2018) showed that the acquisition of abstract concepts is influenced by emotional valence, particularly for children who are at a stage where they are learning abstract words (e.g., 40% of a 12 year old's vocabulary is made up of abstract terms (Borghi et al., 2019), see above). This explains, we conjecture, to some degree why sentiment and emotional valence can be inferred from text in natural language processing tasks, but similar to symbol grounding, emotional valence is inferred from the text, not encoded within it.

Taken together, this suggests that the separation of language from emotion in computational models is going to lead to something that is only an approximation of what a model of language meaning should encode and in that way it is similar to the Symbol Grounding Problem. However, emotion is not just another modality like vision through a camera or haptic sensations through a robotic hand; emotion is communicative on its own, albeit with limited (but important) social signals; pre-linguistic in that it helps scaffold the language learning process especially early on, and emotion is later intertwined with cognitive development and linguistic meaning at an abstract level. Dreyer and Pulvermüller (2018) suggests that representing emotion computationally could be done through the motor system, as done in Moro et al. (2020), which may offer a starting point for bringing emotion into computational models of language (instead of the other way around).

5 Open questions

Resolving the Symbol Grounding Problem has seen real progress, in particular with vision (see below), but it is far from completely solved. There are many modalities to be explored beyond vision, and it is unlikely that the research field will arrive at a solution to representing meaning computationally without some kind of representation of an approximation of emotion. Given the implications of the above sections, in this section we discuss the fact that (besides emotion, discussed above) there are modalities besides just vision that need to be grounded into, pose some open questions, and offer some next steps for the research community.

5.1 Modality questions

Some language models do attempt to model language and vision directly to solve language and vision tasks, for example VilBERT (Lu et al., 2019), CLIP (Jia et al., 2021), FLAMINGO (Alayrac et al., 2022), Dalle 2 (Ramesh et al., 2022), and others. These models are impressive compared to our toy example, but recent work has shown that the models do not quite learn a vision-language mapping in a way that, we argue, actually addresses the Symbol Grounding Problem (Parcalabescu et al., 2020, 2021; Marcus et al., 2022). These language and vision models often force the addition of visual information through robust object detection models that do not capture the true grounding of the words; rather the representation of visual perception is represented symbolically by class labels of the object detection model, but most object detection models do not capture words beyond objects (i.e., nouns). Words like *left* or *red* are also concrete words that an object detection (or region detection) model should not ignore. Moreover, Hendricks et al. (2021) explained that the *quality* of the language (i.e., text) highly affects the visual language models' performance, which seems to suggest that a curriculum not unlike Xu et al. (2020), i.e., by using a training regime that learn with simpler examples first (e.g., that refer to visual objects) then move towards more complex and more abstract examples of language use.

Most recent work has focused on vision, thanks in part to datasets that connect language and vision, but vision isn't the only important modality that humans have access to for grounding linguistic meaning (see (Fernandino et al., 2022; Lynott et al., 2019)): and some have explored grounding into other modalities including modalities that sense the external world like olfactory (Kiela et al., 2015) and sound (Thomason et al., 2018), but also "internal" (i.e., within the body) modalities such as haptics (Thomason et al., 2018), proprioperception (Moro and Kennington, 2018) and interoception (Moro et al., 2020) (i.e., affect & emotion) as well as in spoken interaction itself (see (Larsson, 2018)).

Grounding into external modalities requires some kind of sensor (e.g., cameras for vision and microphones for sound), but more challenging is grounding into internal modalities like haptics, proprioperception, and interoception because for those some kind of embodiment (e.g., robot or virtual agent) is required. We do not explore here which might be better for computationally modeling linguistic meaning, but, following our inspiration from child development above, we make an obvious point that children have bodies that house the sensors and internal modalities that they use to interact with objects and people in the world. Embodied cognition is not a new idea, but given the discussion above, embodiment may be a requirement for capturing holistic linguistic meaning computationally (Barsalou, 2008; Johnson, 2008; Bisk et al., 2020) and embodiment is not in disagreement with solving the Symbol Grounding Problem. The model described in Hill et al. (2020) may be a step in the right direction, though it remains unclear what degree of concreteness or abstraction the model is learning.

5.2 Modeling questions

Much of the recent literature uses vectors and tensors (i.e., within language models) to computationally represent meaning (grounded or ungrounded), which are convenient for hardware that can parallelize computation of such representations, but are vectors the right representation for learning and modeling meaning, particularly meaning that addresses the Symbol Grounding Problem? One possible alternative are *cognitive architectures*. Is it time to work with cognitive scientists and apply their cognitive architectures in spoken, person-to-person interactive settings? Developmental robotics as a field have done so to an extent (Cangelosi and Schlesinger, 2015), and if we are coming to similar conclusions that embodiment may be necessary, but at they very least interactive learning and sensors are required, then it may be prudent to bring more cognitive scientists into the discussion, where possible. More related to concreteness and abstractness, roboticists have worked on making robot actions composed aggregates of smaller, more concrete actions, which may have implications for modeling language.

5.3 Philosophical Questions

It is clear that when Firth posited that meanings of words can be found in the company they keep, the "company" that Firth meant was company with other words, and researchers often cite Wittgenstein for *language is use in context* which always assumes that *context* means lexical context with other words, but Wittgenstein (2010) brings up color and shape (1.72-74) and that words refer to objects, which themselves have affordances (1.11), and early on mentions that language use is first in reference to deictic (i.e. pointing) gestures. Could Wittgenstein have meant that *context* is not lexical context, but physical context (or some degree of both)? This is an important question because Firth and Wittgenstein have always been called on to motivate distributional methods of language modeling, but words keep company with more than just other words, including words that are more concrete.

6 Conclusion

In this paper we attempted to re-frame the Symbol Grounding Problem as a problem of modeling and learning word meanings from concrete as well as abstract words. How meaning of concrete words are modeled and learned follows directly from symbol grounding, and more abstract words could be learned distributionally.

We will build on our toy example in a large-scale experiment by learning classifiers that are not specifically tied to any known grouping of words, but rather are bottom-up grouping of concepts that are linked to works that are later "heard" by the training regime (e.g., in a similar way that someone may know that colors group together based on their features, but do not yet know the word *color*).

We will also explore how such trained classifiers could be combined with existing language model architectures like BERT. Recent work by Kennington (2021) showed how extracting coefficients from visually-grounded classifiers could enrich a language model, but the enriching took place only in the language model's embedding layer with the assumption that all words were concrete. We will explore using concreteness ratings as a possible signal to determine whether a word's meaning should come from a grounded model or a language model.

Acknowledgements Thanks to the anonymous reviewers for their very useful feedback.

References

L Alan Sroufe, Byron Egeland, Elizabeth A Carlson, and W Andrew Collins. 2009. *The Development of the Person: The Minnesota Study of Risk and Adaptation from Birth to Adulthood*. Guilford Press.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for Few-Shot learning.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, (59):617–645.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Association for Computational Linguistics*, pages 5185–5198.

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. 2020. Experience grounds language. *arXiv*.

Yuri Bizzoni and Simon Dobnik. Sky + fire = sunset exploring parallels between visually grounded metaphors and image classifiers.

Anna M Borghi, Laura Barca, Ferdinand Binkofski, Cristiano Castelfranchi, Giovanni Pezzulo, and Luca Tummolini. 2019. Words as social tools: Language, sociality and inner grounding in abstract concepts. *Phys. Life Rev.*, 29:120–153.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behav. Res. Methods*, 46(3):904–911.

Angelo Cangelosi and Matthew Schlesinger. 2015. *Developmental robotics: From babies to robots*. MIT press.

Eve V Clark. 2013. *First language acquisition*. Cambridge University Press.

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Pasquale A Della Rosa, Eleonora Catricalà, Gabriella Vigliocco, and Stefano F Cappa. 2010. Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 italian words. *Behav. Res. Methods*, 42(4):1042–1048.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding.

Felix R Dreyer and Friedemann Pulvermüller. 2018. Abstract semantics in the motor system? – an event-related fMRI study on passive reading of semantic word categories carrying abstract emotional and mental meaning. *Cortex*, 100:52–70.

Leonardo Fernandino, Jia-Qing Tong, Lisa L Conant, Colin J Humphries, and Jeffrey R Binder. 2022. Decoding the information structure underlying the neural representation of concepts. *Proc. Natl. Acad. Sci. U. S. A.*, 119(6).

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42(1-3):335–346.

Stevan Harnad. 2017. To cognize is to categorize: Cognition is categorization. In *Handbook of Categorization in Cognitive Science*, pages 21–54.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers.

Aurélie Herbelot. 2013. What is in a text, what isn't, and what this has to do with lexical semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 321–327, Potsdam, Germany. Association for Computational Linguistics.

Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. 2020. Grounded language learning fast and slow.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and Vision-Language representation learning with noisy text supervision.

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

Casey Kennington. 2021. Enriching language models with visually-grounded word vectors and the Lancaster sensorimotor norms. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 148–157, Online. Association for Computational Linguistics.

Douwe Kiela, Luana Bulat, and Stephen Clark. 2015. Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language*

Processing (Volume 2: Short Papers), pages 231–236, Beijing, China. Association for Computational Linguistics.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behav. Res. Methods*, 44(4):978–990.

George Lakoff and Mark Johnson. 2008. *Metaphors We Live By*. University of Chicago Press.

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

Staffan Larsson. 2018. Grounding as a Side-Effect of grounding. *Top. Cogn. Sci.*

John L Locke. 1995. *The Child's Path to Spoken Language*. Harvard University Press.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic visiolinguistic representations for Vision-and-Language tasks.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2019. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behav. Res. Methods*, pages 1–21.

Gary Marcus, Ernest Davis, and Scott Aaronson. 2022. A very preliminary analysis of DALL-E 2.

Claudia Mazzuca, Luisa Lugli, Mariagrazia Benassi, Roberto Nicoletti, and Anna M Borghi. 2018. Abstract, emotional and concrete concepts and the activation of mouth-hand effectors. *PeerJ*, 6:e5987.

David McNeill and Casey Kennington. 2020. Learning word groundings from humans facilitated by robot emotional displays. In *Proceedings of the 21st Annual SIGdial Meeting on Discourse and Dialogue*, Virtual. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2015. Efficient estimation of word representations in vector space. *In International Conference on Learning Representations (ICLR)*.

Daniele Moro, Gerardo Caracas, David McNeill, and Casey Kennington. 2020. 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.

Daniele Moro and Casey Kennington. 2018. Multimodal visual and simulated muscle activations for grounded semantics of hand-related descriptions. In *Proceedings of the 22nd Workshop onthe Semantics and Pragmatics of Dialogue*.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. VALSE: A Task-Independent benchmark for vision and language models centered on linguistic phenomena.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2020. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks.

Marta Ponari, Courtenay Frazier Norbury, and Gabriella Vigliocco. 2018. Acquisition of abstract concepts is influenced by emotional valence. *Dev. Sci.*, 21(2).

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional image generation with CLIP latents.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv*.

Jacqueline Sachs, Barbara Bard, and Marie L Johnson. 1981. Language learning with restricted input: Case studies of two hearing children of deaf parents. *Appl. Psycholinguist.*, 2(01):33–54.

David Schlangen, Sina Zarriess, and Casey Kennington. 2016. 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*, pages 1213–1223.

John R Searle. 1980. Minds, brains, and programs. *Behav. Brain Sci.*, 3(03):417.

Linda Smith and Michael Gasser. 2005. The development of embodied cognition: Six lessons from babies. *Artif. Life*, (11):13–29.

Mariarosaria Taddeo and Luciano Floridi. 2005. Solving the symbol grounding problem: a critical review of fifteen years of research. *J. Exp. Theor. Artif. Intell.*, 17(4):419–445.

Jesse Thomason, Jivko Sinapov, Raymond J Mooney, and Peter Stone. 2018. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 5520–5527. AAAI.

Caterina Villani, Luisa Lugli, Marco Tullio Liuzza, Roberto Nicoletti, and Anna M Borghi. 2021. Sensorimotor and interoceptive dimensions in concrete and abstract concepts. *J. Mem. Lang.*, 116:104173.

Philippe Vincent-Lamarre, Alexandre Blondin Massé, Marcos Lopes, Mélanie Lord, Odile Marcotte, and Stevan Harnad. 2016. The latent structure of dictionaries. *Top. Cogn. Sci.*, 8(3):625–659.

L Wittgenstein. 2010. Philosophische untersuchungen. In *Sprachwissenschaft*, pages 105–111. De Gruyter.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.