Distributional Semantics for Situated Spatial Language? Functional, Geometric and Perceptual Perspectives

John D. Kelleher[1] and Simon Dobnik*[2]

[1]ADAPT Centre, ICE Research Institute, Technological University Dublin, Ireland
[2]Department of Philosophy, Linguistics and Theory of Science, and the Centre for Linguistic Theory and Studies in Probability (CLASP), University of Gothenburg, Sweden
[1]Email id: john.d.kelleher@tudublin.ie
[2]Email id: simon.dobnik@gu.se

July 7, 2022

Abstract

Distributional semantics has been at the core of recent developments in deep learning work for natural language processing. This distributional semantics plus neural processing paradigm has resulted in significant improvements in state of the art results across a large number of tasks, including parsing, text classification, and machine translation. However, there are a number of areas of natural language processing research where this shift in paradigm has not resulted in significant improvements in system performance. One such area is in situated dialogue systems (such as those studied in the field human-robot interaction), and in particular with respect to the processing of spatial references. This chapter examines why this lack of progress has occurred, through a review of existing research on grounding language in perception that is structured around three forms of semantic

*Both authors contributed equally.

information available in situated dialogue: functional, geometric and perceptual. Through this review we identify which aspects of perceptual grounding distributional semantics naturally accommodates and which aspects it does not. Building on this insight we suggest avenues for future work that attempt to integrate distributional and non-distributional information in order to progress research in perceptual grounding of language, and discuss the broader implications of our findings for computational representations of natural language semantics.

1 Introduction

Distributed (continuous dense vector based) representations have a number of semantic and computational advantages for Natural Language Processing (NLP). For example, a distinctive and pervasive characteristic of natural language is the vagueness of many semantic concepts (van Deemter, 2010). Continuous vector based representations, combined with distance based measures of similarity, provide an intuitive semantic space (where the meaning of a concept is understood as a general region rather than as a point) for encoding vagueness (Karlgren and Kanerva, 2021). Furthermore, understood in this way vector based representations also provide a natural basis for automatic generalisation from terms to concepts (Hinton et al., 1986). At the same time, from a computational perspective vector based representations provide both convenient intra-modal interfaces in multi-modal NLP systems and intra-lingual representations in machine translation systems (Kelleher, 2019).

Notwithstanding the fact that the advantages of distributed representations have been known for some time, for many years symbolic (local) representations–either based on logical formalisms or knowledge engineering–were dominant within NLP. However, systems built using these symbolic formalisms often struggle to adequately manage the ambiguity inherent in natural language, and the brittleness of rules based analysis (Gazdar, 1996). Throughout the 1990s the ground work was laid for the adoption of distributed representations within mainstream NLP. Examples of early work in this direction include (Schütze, 1993) and (Lund and Burgess, 1996). A commonality across both of these works is that the vector representation is learned from data through an analysis of lexical co-occurrence within a large-corpus. Today this focus on lexical co-occurrence is often framed as implementing the *distributional hypothesis* (Harris, 1954), popularised by Firth (1957) as the meaning of "a word is characterised by the company it keeps". These early works on inducing vector based representations through

lexical co-occurrence can be seen as the progenitors of more recent work such as *word2vec* (Mikolov et al., 2013) and the Sesame Street models of ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019).

The combination of distributional semantics with distributed representations, in tandem with modern methods for training of large neural language models, is now the dominant approach in NLP. Indeed, the dominance of this distributional+distributed+neural paradigm has begun to invite criticism, in particular with respect to claims attributing "understanding" to these models. For example, Bender and Koller (2020) propose a distinction between form and meaning—defining "(linguistic) meaning to be the relation between linguistic form and communicative intent"—and build on this distinction to argue that "the language modelling task, because it only uses form as training data, cannot in principle lead to learning of meaning". In a similar vein, Bisk et al. (2020) argue that "meaning does not arise from the statistical distribution of words" and argue that NLP as a field must move beyond solely training on massive internet based corpora, to consider aspects of meaning arising from perception, embodiment and social/interpersonal communication.

Sahlgren and Carlsson (2021) provide a counter voice to these criticisms. Although (Sahlgren and Carlsson, 2021) share a scepticism of claims that current language models achieve natural language understanding, they also argue that criticisms of distributional semantics as a basis for ultimately achieving language "understanding" are founded on the erroneous assumption that language meaning is a single and uniform phenomenon, and error that they name "the singleton fallacy". Furthermore, Sahlgren and Carlsson (2021) argue that criticisms of distributional approaches to semantics that are based on a distinction between form and meaning are fundamentally "dualist" and "defeatist": requiring the creation of an "entire human in silico" before allowing the attribution of natural language understanding to a computational system. By contrast, they propose that even if a distinction between form and meaning holds then there must be a linguistic correlate between form and meaning, and distributionalist approaches access this linguistic correlate enabling a system to access meaning: "in the sense that intentions (meanings) have the effects on linguistic signal (form) it will be possible to learn these effects by simply observing the signal" (Sahlgren and Carlsson, 2021, pg.5).

One thing that Bender and Koller (2020), Bisk et al. (2020) and Sahlgren and Carlsson (2021) do agree on is the importance of multimodality as a future research direction for natural language understanding. On this they also agree that in testing the extent to which a distributional approach can attain understanding of language it is

a unnecessary restriction to only consider distributions across textual data, with all three pointing to the success of research on image captioning (e.g. Xu et al. (2015); Lindh et al. (2018); Herdade et al. (2019); Lindh et al. (2020)). Indeed, Sahlgren and Carlsson (2021) go so far as to argue that these systems can claim a "visual understanding" of language, and pose the question of whether there is anything about language that cannot be learned from a distributional analysis of a large (potentially multimodal) corpus?

In our view, an important omission from the current discussion on the extent to which language meaning is learnable via a distributional analysis of multimodal data is the question of situated spatial language. Visually situated dialogue is spoken from a particular point of view within a physical or simulated context, and from a computational semantics perspective provides a useful test-bed for grounding language meaning in vision (Kelleher, 2003; Dobnik, 2009). From a practical perspective a natural domain of application for situated dialogue systems is human-robot interaction (Kelleher and Kruijff, 2005a; Kruijff et al., 2006a; Dobnik and de Graaf, 2017). From a theoretical semantics perspective the resolution of linguistic references to referent's in the visual domain is interesting for a range of reasons. First, it requires intra-modal fusion (identifying that different occurrence of an object within a modality concern the same object, such as different views of the same object) and inter-modal fusion (Kruijff et al., 2006b). Second, is requires the system to have a perceptual memory both to enable intra-modal fusion (see (Kelleher, 2006) on co-reference classes), but also to enable the system to ground references to entities that were previously seen (Kelleher et al., 2005; Kelleher and Dobnik, 2019; Dobnik and Silfversparre, 2021). An inherent part of situated dialogue is spatial language, including references that refer to entities through their location (i.e., locative expressions) such as *the person near the table* (Kelleher and Costello, 2009), or references to contextually defined spatial regions, such as *the front of the room* in the robot command *go to the front of the room* (Hawes et al., 2012).

An error made by many researchers new to the field of situated spatial language is to equate spatial semantics with Euclidean geometry. However, a deeper analysis of the meaning of spatial descriptions reveals that the extent of the spatial regions linguistically described as *near the table* or *front of the room* are not definable via geometry alone. The definition of these regions are also sensitive to functional use, the perspective of the speaker and hearer, and the distribution of objects and the configuration of the enclosing region. Indeed, we believe that it is the oversimplification, by many researchers, of spatial semantics as equating to geometry, that is part of the reason why the

challenge of spatial language is so often overlooked in discussions on linguistic meaning. This is most clearly seen in discussions relating to the ability of image captioning systems to ground language in vision. For example, as noted above, Sahlgren and Carlsson (2021) posit that these systems can claim to have a "visual understanding" of language despite that fact that, as Kelleher and Dobnik (2017) highlight, the "visual understanding" of spatial language attained by many of these systems is fundamentally limited by the fact that CNN visual encoders focus on *what* is in an image, and (deliberately) discard information relating to *where* an object is. Furthermore, although there are recent image captioning systems that explicitly address the challenge of spatial language (see for example (Yang et al., 2019; Yao et al., 2018; Herdade et al., 2019)) all of these approaches are founded on geometric features extracted from an object detector, and essentially reduce spatial semantics to geometry between bounding boxes.

Motivated by Sahlgren and Carlsson (2021) question of *what cannot be learned from a distributional analysis of a large (potentially multimodal) corpus*, and by a desire to highlight situated spatial language as a challenge for natural language understanding that requires more than geometry, in this chapter we examine aspects of the meaning of situated spatial language that might or might not be learnable via a distributional approach. The chapter begins by reviewing previous work on distribution semantics (Section 2), and this if followed by a review of previous work on situated spatial language (Section 3) that is structured around three themes of research: geometric approaches (Section 3.1), perception (Section 3.2), and functional aspects (Section 3.3). Following this review we discuss experimental work on identifying spatial knowledge from distributions of words over contexts (Section 4) and finally how distributional information interacts with representations of other modalities of probabilistic grounded language models (Section 5).

2 Distributional hypothesis

The history of distributional semantics can be traced at least as far back as (Harris, 1954). However, for this work we will limit our review to the most recent generation of computational work within the distributional tradition, beginning with the models of *word2vec* (Mikolov et al., 2013) and *GloVe* (Pennington et al., 2014).

A fundamental distinction on approaches to distributional semantics induced from corpora can be drawn based on the conceptualisation of the context used in the analysis. For example, some approaches, such as Latent Semantic Analysis (LAS) (Deerwester et al., 1990), use

the documents in the corpus to define context and base the analysis on a "term-document" matrix. The focus of these approaches on the occurrences of terms within a document means that they are most naturally applied to document/information retrieval tasks.

By contrast, both word2vec and GloVe define context in terms of a window of words either side of a word. In these approaches each instance of a term in a corpus is taken to define an instance of a context, with the width of the contexts being hyper-parameter of a model and often being truncated by sentence boundaries. Given these instance+window based contexts, word2vec proceeds by training a neural network to make predictions of term co-occurrences within these contexts, either by predicting what terms co-occur with a given term (i.e., the continuous bag-of-words task) or what term co-occurs with a given context (i.e, the skip-gram task). This neural network model is trained by randomly initialising the distributed representation (embedding) for each term and then using the back-propagated errors of the network on the task to learn the term embeddings. GloVe, on the other hand, begins its analysis by constructing a "term-term" matrix where an entry in the matrix records the number of times that the column term occurred within the context window of an instance of the row term. This matrix is constructed by first identifying for each term in the vocabulary all instances of the term in the corpus, applying the context window onto each instances, and then for each term in the vocabulary recording the number of total number of occurrences of these terms across these contexts. A neural network model is then trained to predict the log of the co-occurrence count of two terms given the embeddings of the two terms as input.

Both of these distributional based models of semantics, GloVe and word2vec, are widely used within natural language processing research. However, as is natural with any successful and widely adopted research approach there has been an array of subsequent research that have critically assessed and extended these approaches in different ways. For example, Kacmajor and Kelleher (2020) highlight the distinction between thematic and taxonomic semantic relatedness, and their experiments indicate that although these models are good at capturing thematic relatedness they are limited in their ability to capture taxonomic relatedness. Interestingly, this limitation, however, provides an example of how the potential limitations of a distributional analysis can be addressed through the design of the corpus the analysis is performed on. For example, a number of researchers have explored combining a random walk process over a lexical taxonomy (such as WordNet) to generate a synthetic corpus so that proximity in the corpus reflects proximity within the taxonomy, and then applying a distributional

analysis to the resulting corpus as a method for injecting taxonomic semantics into embeddings (Maldonado et al., 2019; Klubička et al., 2019; Kacmajor et al., 2020).

Another, perhaps better known, critique of the word2vec and GloVe methods is that they generate a single semantic representation (embedding) for a term, the implication being that this single semantic representation cannot do justice to the variety of meanings a term can have across difference contexts of uses. The best-known solutions developed for this drawback (e.g., ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019)) have retained the core concept of a distributional semantics but coupled this with the contextual dynamism offered by language models, in order to generate contextualised embeddings that represent the meaning of a term in the given context.

Finally, the fact that these distributional based semantic models typically use distributed representations has meant that another critique of these methods is that it is difficult to understand what information they are capturing and encoding through the distributional analysis. This has resulted in a growing body of work focused on developing methods for *probing* these representations—e.g., (Salton et al., 2016; Conneau et al., 2018; Forbes et al., 2019; Ettinger, 2020; Nedumpozhimana and Kelleher, 2021; Ilinykh and Dobnik, 2021a,b, 2022; Nedumpozhimana et al., 2022)—in order to understand what information is encoded within them. This body of analytical work has revealed that these distributional based methods can capture and encode information relating to syntactic information (such as parts of speech, chunks, and subject-predicate agreements), semantic information (relating to semantic roles, entity types and relations), and of particular interest for the focus of this chapter although these methods can struggle with task requiring world knowledge, such as pragmatic inference, they are competitive on knowledge induction tasks (e.g., filling in blanks "cats like to chase __"), and they also have some ability with respect to capturing affordances and properties of objects (Rogers et al., 2020).

The abilities of these methods to capture and encode world knowledge, such as the properties of objects and their affordances, suggest that these distributional methods may be suitable for capturing aspects of spatial linguistic meaning that traditional approaches have struggled with. In the next section we will review the literature of spatial language, and following that we will present our analysis of the suitability of distributional methods to progress research on the spatial language semantics.

3 Situated spatial language

Physical sciences have developed ways in which space can be described with a high degree of accuracy, for example by measuring distances and angles. Furthermore, such measures can be represented on a continuous scale of real numbers. However, humans refer to space quite differently: descriptions such as "the chair is to the left of the table" or "turn right at the next crossroad" refer to discrete units such as points, regions and volumes and require knowledge about how the objects related interact with each other. We can also take different spatial perspectives which are frequently not explicitly described. For example, the same object can be to the left or to the right of another object depending on the viewpoint taken. Human spatial descriptions bridge perceptual and conceptual domains and are notoriously vague. Consequently, they need to be evaluated relative to each other and relative to the perceptual (visual salience, cf. (Kelleher et al., 2005)) and linguistic (discourse salience, cf. (Brennan and Clark, 1996)) contexts in which they occur and which changes as the interaction unfolds.

Spatial descriptions have been studied extensively in linguistics, computational linguistics, psychology, computer science and geo-information science. Classical surveys of their semantics can be found in Herskovits (1986), Miller and Johnson-Laird (1976) and Talmy (2000). There are different categories of spatial descriptions which have slightly different properties. *Locative descriptions* describe locations of objects and individuals in scenes, for example "the chair is to the left of the table" or "the chair is between the table and the sofa". The object whose locations is being described is normally described as *target* while the other object whose location must be previously known is called the *landmark*. The spatial (locating) relationship between the target object and the landmark is frequently described using spatial terms that involve prepositions (e.g., 'X on Y', 'X in front of Y'). In English there are approximately 80 spatial prepositional descriptions in common use (Landau, 1996), and a distinction is made between those that primarily indicate a topological (i.e., contact, support, and so on) or proximal relationship (e.g., *at*, *on*, *in*, *near*, and so on), and those that indicate a directional or projective relationship (e.g., *behind*, *above*, *in front of*, and so on) (Kelleher and Kruijff, 2006). Spatial descriptions can also be describing actions in which case the relations involve verbs, for example "a skater is jumping over a fire hydrant". In this case the verb is further describing the nature of relation between the objects. Hence, the previous description of action is parallel to "a skater over a fire hydrant". *Route descriptions* and *route instructions* such as "follow this road until the next crossroad and then turn left" are another class of

spatial descriptions where there is no explicit landmark object, since this is implicitly assumed to be the agent following the instruction.

Although the challenges for their modelling are well defined descriptively, to date there has been no unified computational model to represent their meaning. Early attempts in this direction were based on first order logic representations, see for example (Winograd, 1976) and (Miller and Johnson-Laird, 1976). However, these only represent the conceptual knowledge and ignore how this knowledge maps to the environment. Since these early works there has been a significant amount of work on spatial semantics that draw on the environmental context (often modelled through geometric models), perceptual factors and cues (such as visual occlusion and perspective), and world knowledge (capturing functional roles and affordances).

3.1 The geometry of the environment

From a computational semantics perspective the contextual environment of a spatial reference is often represented through a two-dimensional or three-dimensional coordinate frame in which we can represent objects and angles and distances between them. Psycho-linguistic research has found that people decide whether a spatial relation between a target object and an landmark object applies by anchoring a *spatial template* on the landmark that defines regions of differing degrees acceptability around the landmark where a given relation holds, see e.g. (Gapp, 1994; Logan and Sadler, 1996). Often these experiments were deliberately designed to exclude confounding factors, such as object roles and affordances and so forth. For example, in the experiments reported in Logan and Sadler (1996) the visual stimuli were 7-by-7 grid marked white regions with an O in the centre cell of the grid and an X in one of the other grid cells, and the accompanying linguistic descriptions of the form *The X is near to the O*. Given paired visual and linguistic stimuli subjects were asked to rate the acceptability of the description relative to the visual ground. Using this experimental setup, Logan and Sadler (1996) studied the semantics of the following prepositions: *above*, *below*, *over*, *under*, *left of*, *right of*, *next to*, *away from*, *near to*, and *far from*. The analysis of their results revealed that patterns of acceptability varied across these prepositions (i.e., each preposition has its own spatial template), but that there was similarity between groups of prepositions (for example, *next to* and *near to* have similar meanings). Table 1 lists the spatial template for *near to* as reported in Logan and Sadler (1996). The analysis of these spatial templates revealed a number of parameters relating to the extent of different regions of acceptability for given prepositions,

1.74	1.90	2.84	3.16	2.34	1.81	2.13
2.61	3.84	4.66	4.97	4.90	3.56	3.26
4.06	5.56	7.55	7.97	7.29	4.80	3.91
4.47	5.91	8.52	**O**	7.90	6.13	4.46
3.47	4.81	6.94	7.56	7.31	5.59	3.63
3.25	4.03	4.50	4.78	4.41	3.47	3.10
1.84	2.23	2.03	3.06	2.53	2.13	2.00

Table 1: 7-by-7 cell grid with mean goodness ratings for the relation *the X is near to the O* as a function of the position occupied by X as reported in Logan and Sadler (1996).

such as the 90° maximum angle of deviation from the canonical direction vector for the acceptability of a projective prepositions (such as above *above*). Later experimental work explored how the shape and extent of spatial templates are affected by other contextual factors. For example, Costello and Kelleher (2006) explored how the presence and location of other (*distractor*) objects affected on the semantics of proximity (see Figure 1); Kelleher et al. (2009) investigated how topological distinctions, such as contact or overlap, affected acceptability judgements and distinctions between *at*, *on*, *in*, and *near*. Building on these experimental results a number of geometrically defined models of spatial semantics have been proposed, including (Regier and Carlson, 2001; Kelleher and Kruijff, 2005b; Kelleher et al., 2006; Kelleher and van Genabith, 2006). Several of these models have been implemented in robotic systems and have been used to ground the robot's interpretation of spatial commands in its sensor data, e.g. Gorniak and Roy (2004); Brenner et al. (2007); Schütte et al. (2017); Dobnik (2009).

3.2 Perception, cognition, and perspective

Moving beyond the geometric relationships between entities in an environment, spatial semantics is also dependent on a range of factors that arise through embodiment, cognition, perception and perspective. Ullman (1980, 1984) argues against a 'direct perception' basis of visual perception and cognition, and instead proposes that the base representation provided by the visual perception module likely does not contain information about spatial relations but rather this derived through *visual routines*. This view is supported by psycho-linguistic evidence that suggests that spatial relations are not pre-attentively available, (Treisman and Gormican, 1988), and that their perception requires attention (Logan, 1994, 1995).

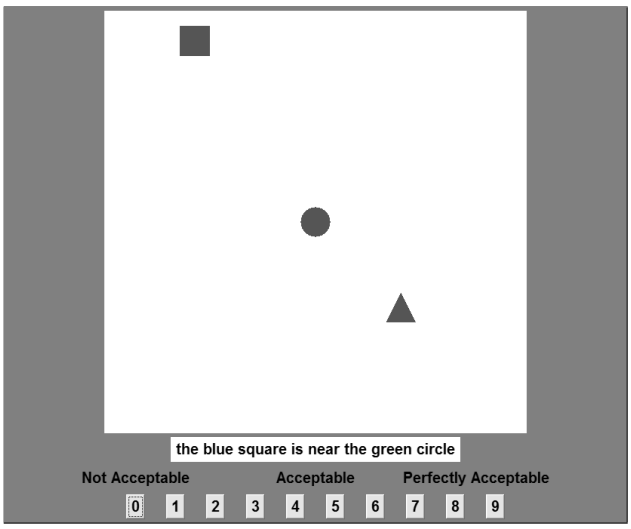


Figure 1: A example trial stimulus from the experiment reported in (Costello and Kelleher, 2006)

It is certainly the case that perception, and cognitive processes relating to perception (e.g., perceptual phenomena such as visual occlusion, visual attention, perceptual memory), affect the processing of spatial language. For example, eye-tracking research has found that visual context influences spoken word recognition and syntactic processing, even during the earliest stages of language processing (Tanenhaus et al., 1995). Also, neuro-physiological studies have suggested two distinct visual neural pathways: the *ventral* pathway processes extract information relating to *what* is in a scene, whereas the *dorsal* pathway—the *where*/*how*—extracts the information necessary to inform the grasping and manipulation of objects, such as size, orientation and location (Mishkin et al., 1983). Landau and Jackendoff (1993) use this neuro-physiological distinction to build an explanation of why language only distinguishes a relatively small number of distinct spatial relations as compared with the human capacity to encode and differentiate a very large number of faces and object types, and further suggest that *dorsal*, or *where*, pathway is the basis for the processing of spatial semantics (the reasoning being here that this pathway only provides a limited information about the detailed geometry of a visual scene and this limited information basis restricts the number of spatial distinctions encoded by language). Extending perceptual processing to include attentional mechanisms and constraints, the *attention vector sum model* proposed by Regier and Carlson (2001) can be understood as extending the *spatial template* paradigm to accommodate the role of visual attention in the processing of spatial templates. Furthermore,

the interaction between visual and linguistic attention and perceptual memory has also been studied and modelled (Kelleher and van Genabith, 2004; Kelleher, 2006; Kelleher et al., 2005; Kelleher and Dobnik, 2019).

Another important cognitive factor that impinges on the semantics of spatial description is the perspective that the speaker and hearer assume is relevant for the spatial description. This is most obvious in the interpretation of directional (or projective) descriptions, such as "to the left of". These require a model of perspective which includes a viewpoint parameter (Maillat, 2003). In literature on spatial semantics, this model of perspective is often termed the *frame of reference*, and computationally modelled as consisting of six half-line axes with their origin at the landmark. In English these half-line axes are usually labelled *front*, *back*, *right*, *left*, *above*, *below*. Three different frames of reference are distinguished in most European languages (Levinson, 1996):

- *intrinsic/object centred*: the orientation of the co-ordinate system is aligned with the landmark object; e.g., the front direction is aligned with the direction the front of the landmark object is pointing towards.
- *relative/viewer centred/deictic*: assumes an ego-centric (observer's) perspective on the scene (typically that of the speaker)—distinct from the orientation of the objects being described—and the orientation of the co-ordinate system is aligned with the observer's view (i.e., front being defined as 'in front of' the speaker)
- *absolute/environment centred/allocentric*: the orientation of the co-ordinate system is aligned with salient properties of the environment; e.g., the above-below axes is aligned with gravity.

The viewpoint may be defined linguistically "from your view" or "from there" but it is frequently left out. This can be either inferred from the perceptual context (if only one interpretation is possible), object affordances ("a person behind the counter") or dynamics of perceptual and linguistic interaction in which case it is negotiated and aligned between conversational partners (Dobnik et al., 2014, 2015, 2020). There is a significant body of experimental work that has examined how ambiguity in relation to intended frame of reference affects cognitive processing and interpretation of a spatial description (Carlson-Radvansky and Irwin, 1993, 1994; Carlson-Radvansky and Logan, 1997; Kelleher and Costello, 2005; Li et al., 2011; Schultheis and Carlson, 2017). One of the findings from this work being that ambiguity with respect to the intended perspective (reference frame) of the speaker can result in multiple spatial templates being activated during the interpretation

and competing and interacting to form a new context specific spatial template (Schultheis and Carlson, 2018).

The examples of directional/projective prepositions is also useful in terms of motivating work on the effect of perceptual phenomena, such as visual occlusion, on spatial semantics. For example, Jackendoff and Landau (1991) argue that visual occlusion likely impacts on the semantics of projective prepositions. Similarly the analysis of Vandeloise (1991) on the semantic distinctions between the French prepositions "devant/derriere" (approximately "before" and "behind") states that in many instances the distinction is based on the presence of absence of object occlusion. Furthermore, Kelleher et al. (2011) report psycho-linguistic experiments that demonstrate the effect of perceptual occlusion on the semantics of *in front of*, and that the predictive performance of the attention vector sum model of Regier and Carlson (2001) can be improved by integrating occlusion as a factor into the model.

The fact that on the one hand spatial relations are not pre-attentively available (i.e., the extracting and processing of these relations are triggered by language), and on the other that there is a variety of perceptual and cognitive factors that affect spatial semantics, raises the question of whether the identification and processing of spatial relationships is primarily driven by outputs of the visual system or by language. To paraphrase Talmy (1983): does language structure the perception of space, or does perception of space structure the processing of language? Allowing some stretching and tolerance in the discussion, this question of the primacy of language or perception for linguistic spatial semantics echoes the discussion we made in the introduction relating the proposition from Sahlgren and Carlsson (2021) that even if a distinction holds between form and meaning there must be a linguistic correlate between them, and distributionalist approaches can access this correlate. Broadly in this vein, Jackendoff (1985) has argued that grammatical structure offers an important source of evidence for understanding cognition, and uses the semantics of spatial expressions as a core exemplar in his analysis. By contrast, Lakoff and Johnson (1980) put space and spatial perception as a core building block of conceptual structure, and hence the extensive use of *spatial metaphor* in language. However, in more recent years there has been a growing awareness of the role of embodiment, the grounding of semantics in action, and the related concepts of object function and affordances. We will review work on these concepts in relation to spatial language in the following section.

3.3 Embodiment, world knowledge, object function and affordances

Noë (2004) argues that perception depends on the capacity for action: "Perceptual experience acquires content thanks to our possession of bodily skills. What we perceive is determined by what we do (or what we know how to do)." Although spatial linguistic semantics is not a central concern for Noë, his *enactive* approach to meaning aligns with a range of work on embodiment and spatial meaning. For example, Barsalou et al. (1999) sets out a perceptual theory of knowledge where the brain captures patterns of sensory-motor activations during perceptual experience, then through selective attention schematic representations of these sensory-motor patterns are stored in memory to form perceptual symbols. These perceptual symbols then become organised into *simulators*. Words then acquire meaning through associations with the simulators for the entities and events to which they refer. Of particular relevance to this discussion, Barsalou argues that spatial relations (and spatial terms) can be understood as simulators:

> "During the perception of a balloon above a cloud, for example, selective attention focuses on the occupied regions of space, filtering out the entities in them. As a result, a schematic representation of above develops that contains two schematic regions of space within one of several possible reference frames .... Following the similar extraction of information on other occasions, a simulator develops that can render many different *above* relations." (Barsalou et al., 1999, p.593)

Roy (2005) adopts a similar perspective to Barsalou and Noë, and proposes *semiotic schemas* as a framework for grounding language in action and perception. In this framework a schema is a structured network of belief that is grounded in an agent's environment through a causal-predictive cycle of action and perception. Furthermore, the basis of language production and comprehension involves operations on schemas. The form of schema most relevant to spatial descriptions, such as "the ball is in the cup", are *situation schemas* which might encode beliefs that the ball and cup are in *contact* and that the cup *contains* the ball. These beliefs are connected to perception and action by, for example if the ball and cup are in contact then the agent can predict that if they look at the cup they will likely see the ball in the vicinity, similarly the belief relating to cup containing the ball can be linked to action and perception by the fact that if the cup (the container) is moved the ball should also be perceived to move.

Building on the theoretical and empirical work relating spatial semantics to embodiment, action and perception, Coventry and Garrod (2004) propose a *functional geometric framework* for treatment of the semantics of spatial prepositions. A core element of the framework is a recognition and foregrounding of the affect of an object's functional role or affordances in the processing of some spatial prepositions. For example, Coventry and Garrod (2004) report experiments that examined differences in how functional roles and geometry affected the interpretation of the prepositions *over*, *under*, *above*, and *below*. In one of these experiment one object (an umbrella) had the function of protecting another object (a person) from falling rain. In this experiment the visual stimuli used in a trial consisted of drawing of a man holding an open umbrella and in some images rain is falling and in others there is no rain. The geometric relationship between the man and the umbrella varied across the images. In one set of images the umbrella was in the normal position of use (aligned with the gravitational plane and above the man); in a second set of images the umbrella was rotated 45° of vertical so that it was diagonally in front of and above the man; and, in a third set of images the umbrella was rotated 90° of vertical so that is was pointing along the horizontal axis in front of the man. Across these three sets of images there was no rain falling, or rain was falling and either falling onto the umbrella or the man. Each trial consisted of presenting one of the images to the subject along with a statement of the form *The umbrella is over the man* or *The umbrella is above the man*, and asking the subjects to score the appropriateness of the description. The analysis revealed that the variation in the geometric relationship between the man and the umbrella has the expected affect: rotating the umbrella away from the gravitational plane reduced the appropriateness of both *over* and *above*. The analysis also revealed, however, that the umbrella's functional role of protection also affected the appropriateness ratings. Scenes where the umbrella blocked the rain received higher scores than the corresponding umbrella and man configuration scenes where the umbrella did not block the rain. These studies also found that *over* and *under* were more influenced by functional roles as compared with *above* and *below*, and conversely that *above* and *below* were more sensitive to geometric factors. Overall the functional geometric framework can be understood as calling for a multi-factorial approach to spatial semantics, as Coventry and Garrod argue the "functional geometric framework, therefore, requires geometric routines, extra-geometric routines that capture dynamic-kinematic relations between objects, and stored representations that reflect stereotypical functional relations between objects" (Coventry and Garrod, 2004, p.127). The breadth of the fac-

tors considered in the functional-geometric framework together with the perceptual, cognitive, and embodied perspective reviewed above demonstrates the complexity and richness of situated spatial language as a test case for linguistic semantics.

4 What can a distributional analysis reveal about spatial language?

Our review of the literature on spatial language highlighted three major sources of information that inform spatial language use: the geometry of the environment; perception, cognition and perspective; and embodiment, world knowledge, object function and affordances. It is relatively easy to programme a computer to implement geometric models. Consequently, the majority of previous computational work on modelling spatial language has focused on creating geometric models of the environment based on sensor readings and implementing semantic models for spatial terms based on functions that define spatial templates within these geometric models. By contrast relatively little computational work has been done on modelling functional factors affecting spatial language use. This is primarily because, compared with geometric considerations, it is not obvious how to capture and integrate functional knowledge into computational models. However, here we argue that distributional analysis may be a useful approach to capture this functional relationships between objects. As we noted in the introduction, probing research on distributed representations learned through distributional analysis of text reveals that these embeddings do encode information with respect to affordances and properties of objects. In this section we explore the potential for distributional analysis to capture these functional based relationships between objects.

4.1 Distributional analysis of image captions reveals the functional and geometric bias of spatial prepositions

(Dobnik and Kelleher, 2013) describes two experiments where the *functional* and *geometric* bias of spatial relations is identified automatically from word distributions in the corpus text and the results are compared with the results of the psycholinguistics studies Coventry and Garrod (2004). The approach uses the textual parts of the IAPR TC-12 Benchmark corpus (Grubinger et al., 2006) and the 8K Image Flickr dataset (Rashtchian et al., 2010). It is important for the study

that the descriptions are extracted from vision and language corpora so that the descriptions are constrained by the visual scene. A preliminary study was done on relations extracted from the British National Corpus (BNC) where the results were not clear cut since there were also frequent non-spatial metaphorical uses of the same relations, for example "in/over three days". The descriptions from both corpora are parsed for dependencies and then rules are manually constructed to normalise words and extract semantic representations of the form Relation(Target, Landmark). The reported results depend on the quality of the dependency parsing and the coverage of the extraction patterns.

The first experiment examines whether the strength of association between the target and landmark objects as captured in a language model corresponds to how strongly the objects are functionally related. The intuition underpinning the first experiment was that functional constraints are stronger than geometric ones, because they are focusing on particular dynamic kinematic routines between objects that are otherwise geometrically relatable. To measure the strength of the association between objects and relations Dobnik and Kelleher (2013) use Log Likelihood Ratio (Dunning, 1993),

$$log\lambda = log\frac{L(H_1)}{L(H2)} \tag{1}$$

which is a log ratio of the likelihood of the hypothesis H_1 over the hypothesis H_2 where H_1 is that the words in a bigram w_1w_2 are independent and H_2 that they are dependent. It therefore tells us how many times more likely H_2 is compared to H_1 and $-2log\lambda$ approximates the χ^2 distribution. The results indicate that the most likely relation for a target and landmark pair "boy-shirt" is "in" and then "with". The most likely relations for the landmark "umbrella" and some other target (e.g. people, boy, table, child, sculpture...) are either "with" and "under".

The second experiment reported in Dobnik and Kelleher (2013) was motivated by the observation that different relations have different selectional properties that can be related and this should be reflected in an information theoretic measure such as *entropy*. If some relations are functionally biased towards particular target-landmark pairs then this should be reflected in a lower entropy over the target-landmark pairs that occur with that relation in the corpus. Note that entropy measures uncertainty, via the distribution across possibilities, and so an entropy based analysis of a corpus naturally aligns with a distributional perspective on language. For example, in the experimental studies such as (Coventry and Garrod, 2004) "in", "on" and "over" have been identified as being influenced by the functional component and

these relations are predicted by the corpus study to have a low entropy of target-landmark pairs compared to their corresponding geometrically biased variants such as "above". (Dobnik and Kelleher, 2013) did find that functionally influenced prepositions, such as "over", do tend to have lower entropy over their target-landmark pairs as compared with non-functionally sensitive counterpart, such as "above".

Dobnik and Kelleher (2014) present two additional experiments based on the same datasets. First, the target and landmark objects that occur with a particular relation are clustered in conceptual categories. The conceptual categories of targets and landmarks occurring with each relation are determined using the synsets of WordNet and the *class-labelling algorithm* (Widdows, 2003) which given a list of words finds a hypernym that subsumes as many words in the list as possible and is applied recursively until all the words are exhausted. The extracted conceptual categories represent semantic classes of target and landmark objects that occur with each relation. Then, based on these clusters, patterns of relations relating different conceptual categories are automatically extracted. Following the previous intuition that functional relations constrain more specific objects it is expected that functional relations will give rise to more patterns as well as these patterns containing more specific classes of objects, and hence more conceptual categories. Tuples ⟨target, relation, landmark⟩ are rewritten as ⟨target-class, relation, landmark⟩ such as (i) travel.v.01 over object.n.01 (9/713), bridge.n.01 over object.n.01 (23/713), bridge.n.01 over body of water.n.01 (42/713) and (ii) person.n.01 under tree.n.01 (7/213), shirt.n.01 under sweater.n.01 (8/213), person.n.01 under body of water.n.01 (11/213), person.n.01 under artifact.n.01 (13/213). The numbers in the brackets indicate the number of times an example following that pattern occurs in the dataset out of the total number of times a relation is found. The automatically extracted patterns reveal different situations of objects being related reflecting human conceptualisation of objects and space which require quite distinct geometric arrangements. In fact, they are similar to examples of spatial relations reported in linguistic literature such as (Herskovits, 1986; Levinson, 2003).

Returning to the question of assessing the functional sensitivity of different spatial relations, the question of how general and specific are the relational patterns for a given preposition arises: recall the intuition that the more functionally sensitive the semantics of a spatial preposition the more specific the objects it encodes relationships between? To estimate the specificity of the extracted patterns, and hence the specificity (functional sensitivity) of the preposition that occurs within the patterns, Dobnik and Kelleher (2014) calculated the

following measures: (i) the average depth of the target and landmark synset hypernyms, (ii) the number of patterns created, and (iii) the entropy of examples of these patterns in the dataset. Their results were that the ranking of spatial prepositions across these three measures in terms of specificity of the patterns the prepositions occur within had high agreement. Furthermore, in these rankings spatial preposition that are known to be functionally sensitive—based on experimental work such as (Coventry and Garrod, 2004)—were found to be more specific in terms of the patterns they occurred in.

What the experiments in (Dobnik and Kelleher, 2013) and (Dobnik and Kelleher, 2014) demonstrate is that distributional analysis of corpora can reveal information relating to the functional semantics of spatial prepositions.

4.2 Neural language models attain different perplexity when generating functional and geometric relations

The previous method extracts targets and landmarks for a given relation independently to form patterns of relations between them. However, during interaction a describer chooses one of the objects they would like to refer to and then a particular landmark conditioned on the target: this has to be functionally associated with the target and visually and linguistically salient in the interaction. The interpreter exploits this salience of the landmark to locate a suitable target which means that the conditioning is reversed here: potential targets are identified by being conditioned on the landmark. Since a relation describes a particular functional and geometric interaction between target and landmark it is conditioned by both. (Dobnik et al., 2018) examines how such relations are captured in a *neural language model* (Bengio et al., 2003; Mikolov et al., 2010; Salton et al., 2017) and also attempts to identify their functional geometric bias by examining language model *perplexity*.

A neural language model estimates the probabilities of a sequence of words by optimising the parameters of the neural architecture based on *cross-entropy loss*. The loss of a neural language model comparing the predicted words with the ground truth is the average *surprisal* over that batch of data. From the loss a measure of *perplexity* can be defined for a particular sequence of words T as follows:

$$Perplexity(S, P) = 2^{E_S[-log_2(P(w_{1:T}))]} \tag{2}$$

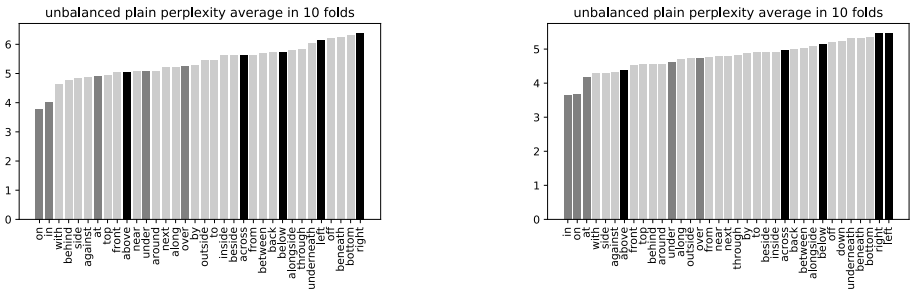
Perplexity therefore indicates a measure of fit of the language model with a sample sequence. Returning to spatial relations, if descriptions

involving functional spatial relations are more selective in choice of their vocabulary (target, landmark and relation) then such sequences are more predictable in the dataset which means that this will result in lower perplexity of the language model.

Dobnik et al. (2018) use the Visual Genome corpus (Krishna et al., 2017) for their experiments. The Visual Genome corpus is a crowdsource annotated corpus of 108K images including *relationships* between objects identified in bounding boxes, for example "cup on table", "girl holding on to bear" and "woman standing on snow". The data is split into 10-folds and a neural language model is trained and validated over these sequences and average results are reported. Dobnik et al. (2018) extracted static spatial relations from these patterns using a dictionary of terms reported in (Landau, 1996; Herskovits, 1986). Other pre-processing steps include re-writing composite spatial relations as single tokens ("jumping_over" and "to the left of", "to left of" → "left") and some additional text normalisation and selection was performed to reduce unwanted variation in the text. The descriptions were then structured into sequences following the pattern <s> `target relation landmark` </s>. Once the pre-processing was completed the data was split into 10-folds and a neural language model was trained and validated over these sequences and average results calculated.

Figure 2 shows the average perplexities of descriptions containing particular spatial relations. Dark grey and black identify relations that are known to have a functional or geometric bias respectively. Descriptions of the same bias cluster together. Furthermore, descriptions containing relations with a functional bias are more predictable by a language model than those that have a geometric bias. It therefore follows that in the latter relations a geometric grounding plays a stronger role in order for them to be interpretable. Overall, these experiments demonstrate that language models learn specific contexts for both functionally and geometrically-biased spatial relations.

In a follow up experiment Ghanimifard and Dobnik (2019a) evaluate these contexts more closely using the same method. We have seen that spatial relations are similar or dissimilar depending on what kind of context target and landmark objects they occur with. If two relations occur with similar targets and landmarks, then projecting a relation in the context of the second relation will lead to a low average surprisal of the language model measured by its perplexity. This way all relations can be compared with each other: descriptions containing a particular relation are first collected and then the relation is swapped by another relation and an average perplexity of the language model over this collection of descriptions is calculated. This is repeated for



(a) test-set (b) training set

Figure 2: Mean perplexities of spatial descriptions of LM1 (dark grey: functionally biased, black: geometrically biased relations).

all relations. The average perplexity values over the relations can then be represented as values of a vector defined by the contexts. Using this methodology Ghanimifard and Dobnik (2019a) analysed the swapability of relations, in terms of the relative increase of perplexity when a relation is swapped. Figure 3 shows the perplexity vectors for 28 relations over 26 contexts. K-means clustering of these contextual vectors identifies clusters of similar vectors and relations that have similar selectional biases.

These perplexity vectors for spatial relations appear to be so strongly discriminative of the relation so that they can be used in several common semantic reasoning tasks. For example, they were tested in the odd-one-out task where relations are grouped on the basis of the geometric criteria, e.g. axes "left" and "right", containment "in" and "out" and proximity "near". To these pairs an odd word is added from another pair that the model must identify. The results indicate that the perplexity vector of a language model trained on textual information can discriminate words by the geometric criteria without ever seeing the scenes that descriptions are referring to. This is because functional and geometric knowledge are complementary and bias to one particular knowledge means less bias to the other. Overall, the work demonstrates that although spatial relations have different selectional requirements there are similarities between their functional and geometric bias. Furthermore, knowing the functional bias one is able to predict also the geometric bias since they are in complementary distribution. This way language models might be capturing some grounding information about the world without ever experiencing the world.

If functional and geometric bias are complementary then differences for different spatial relations should also be observable in the

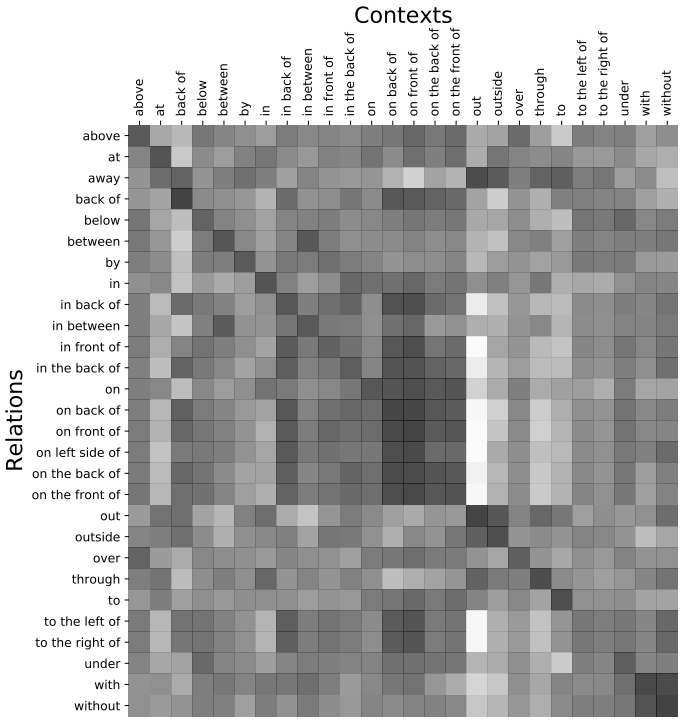


Figure 3: A matrix of perplexity vectors for 28 spatial relations and 26 contexts.

geometric domain. The literature on spatial cognition gives a plenty of examples of situations where particular relation would be chosen by a speaker but where objects are displaced from the typical location where we would expected them to be based on the geometric considerations alone. For example, "the umbrella is over a man" still holds when this is held horizontally providing that it is protecting the person from the rain and "apples are in a bowl" is acceptable where the volumes of some apples are not contained in the volume of the bowl as long as the bowl is constraining the movement of the apples so that they do not fall of the heap of the apples that are geometrically contained in the bowl. This suggest that due to the effect of the functional interactions between objects their locations represented as bounding boxes are expected to be more variable and diverging from the geometric axes, thus *where* they are. However, such relations are more restrictive in terms of *what* objects they relate to as shown in the previous discussion. The opposite is expected for geometrically biased relations.

(Dobnik and Ghanimifard, 2020) tests this hypothesis on the Visual Genome corpus (Krishna et al., 2017) using a similar extracting and pre-processing method to the one described earlier, this time also including descriptions of actions that refer to dynamic relations between objects. To calculate the relation between bounding boxes of target and landmark for a particular relation *dense vectors* with dimensions $[x, y, d]$ are created, where x and y represent directions between two points p_1 representing the target and p_2 representing the landmark in a 2-dimensional space and d is the Euclidean distance between these two points. The distance is assigned a negative prefix if p_2 from the target is also a point in the landmark. Otherwise, the distance is a positive value. The vectors are inspired by the Attentional Vector Sum (AVS) model (Regier and Carlson, 2001) in terms of capturing directions and by the spatial template model (Logan and Sadler, 1996) in terms of capturing distances. The two kinds of representations are also motivated by the difference between *projective* ("to the left of") and *topological relations* ("near"). While the former are sensitive to distance and direction, the latter are only sensitive to distance. An image is segmented into a grid of 7×7 locations and for each location representing target and landmark such a vector is calculated. Vectors are then collected across images for each instance of that relation between objects and finally averaged into a single vector. Using a similar method spatial templates from (Logan and Sadler, 1996) are also converted to single comparable vectors. A cosine similarity between these vectors indicates that vectors extracted from the Visual Genome are similar to those vectors from the spatial template experiments which validates the approach.

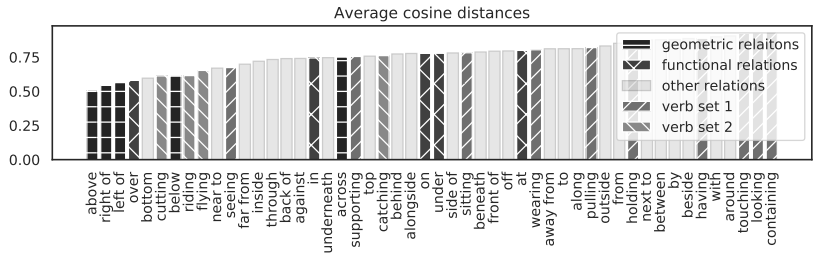


Figure 4: The average cosine distance of dense vectors $[x, y, d]$ from the average dense vector per spatial relation.

In order to estimate the variation of target and landmark objects that is indicative of functional and geometric bias a deviation of individual vectors from the average vector for a particular relation is determined using cosine distance as shown in Figure 4. An average cosine distance close to 0 indicates a stronger central tendency and therefore a geometric bias. Functionally and geometrically biased relations reported in the literature are indicated in brown and blue respectively. Verb set 1 are those verbs reported in (Collell et al., 2018) where the location of the landmark is predicted the least from the y dimension (e.g. "see") and Verb set 2 are those where the location of the target is most strongly predicted from the y dimension (e.g. "flying"). Figure 4 demonstrates that the dense vectors for the geometrically-biased relations tend to deviate from the mean dense vectors less than those for functionally-biased relations and the same trend is observed with the two groups of verbs. The clustering of cosine similarities for individual relations reveals different contexts of object interactions resembling different conceptualisations of (Herskovits, 1986) and that there may be an overlap of vectors between geometric and functional locations where both of these constraints are satisfied.

The preceding discussion demonstrates that an important part of semantics of spatial relations which also include verbs is captured by distributed representations that are estimated from word contexts. The semantics that distributed representations capture are those that refer to functional interactions between objects, the dynamic kinematic routines that reflect human take on the world and are therefore part of the common-sense semantic knowledge about *what* objects are interacting. Since objects and events are taking place in physical space different relations also refer to *where* these objects are, how they are grounded in the geometric representation of space. Different relations are biased to function or geometry differently and so the bias can be evaluated on a scale rather than being fixed for two classes of relations.

Importantly, since the two dimensions of meaning are complementary to each other, a bias to one is also reflected in the other and therefore predictions of one also has implications on the prediction of the other. Furthermore, the experimental study in (Dobnik and Åstbom, 2017) suggests that this bias is not fixed even for individual relations and that contextual factors might affect it.

5 Grounded language models

In the previous section we have examined the contribution of semantic information for different kind of linguistic descriptions but how is this information captured as representations of probabilistic neural language models? We call language models that contain semantic information from word and perceptual contexts *probabilistic grounded language models*. Descriptions contain sequences of words that might be grounded in different modalities to a different degree as well as there are dependencies between words or structures in these descriptions which define the *compositionality* of natural language. (Ghanimifard and Dobnik, 2017) examines to what degree a neural language model is able to learn compositionality of complex spatial descriptions that contain one or more spatial relations, connectives such as "and", "either" and "or", the modifier "not" and other (distractor) words such as functional words which are expected to have no perceptual but only linguistic grounding in the word contexts. If meaning of linguistic representations is compositional as suggested by formal semantics (Montague, 1974; Blackburn and Bos, 2005) which requires that every composed semantic representation has an interpretation relative to a model, then in the case of grounded language models we should observe that the predictions of sequences of words will map to the underlying perceptual representations.

Ghanimifard and Dobnik (2017) generate an artificial dataset of descriptions over locations from the acceptability judgements reported in (Logan and Sadler, 1996), for the prepositions *above*, *below*, *over*, *under*, *left of*, *right of*, *next to*, *away from*, *near to* and *far from*, and the connectives described above. Locations are the 48 locations the target object can occupy in the 7×7 grid used in the experiment, see Table 1 (the 49th location is the centre where the landmark is located). Examples are created for individual expressions. For the grounded language model an LSTM is used as encoder and decoder. The input is an embedding vector learned from the one-hot encoded words which is concatenated with the location of the target and the output is a prediction of the following word. For example, assuming the target object is located in the top left quadrant of the grid, if the model is

provided the input embedding for the sequence *left of and* then the correct next word prediction would be *above*. The model is evaluated on held-out sequences of composed descriptions where individual word tokens have been seen during training but their compositions are novel to the model. The next word probabilities over the vocabulary returned by a model can be thought of as acceptability judgements of words for that particular input location and can be aggregated to create spatial templates. The spatial templates generated from the data of a grounded neural language model on new composed phrases can then be compared with the original spatial templates from which the training data has been generated. Correlation with Spearman's ρ indicates that overall there is a high correlation between the predicated and the ground truth spatial templates when evaluated on the same dataset as the training one. Similarly the model is also able to predict simple descriptions containing a single relation while only seeing composed phrases during training. Effectively, the model has learned how to decompose phrases in terms of perceptual grounding.

These results indicate that grounded neural language models can capture very well semantic information from word and perceptual contexts in the form of *grounded dense word embeddings*. However, it is important to note that not all language models ground the meanings of spatial relations, indeed during generation (for example in an image captioning scenario) language models often generate ungrounded relations which are also known as *hallucinations* of language models.

For example, Ghanimifard and Dobnik (2018) examine the adaptive attention in a pre-trained model of (Lu et al., 2017). They start by extracting spatial descriptions and their corresponding images from the MS-COCO test corpus and investigating where the model directed its attention to when processing an image and a spatial description. Note that in this model the attention is explicitly modelled as complementary between the attention on the image features and attention from the language model. A comparison of the POS-tags of words that receive visual attention reveals that this is primarily focused on those tags that make up noun-phrases, e.g. NUM, NOUN, ADJ, DET, while VERB receives moderate attention and ADV(erbs) and ADP(ositions) are among the least attend POS-tags, only above PRON and PRT. Average adaptive attention on words referring to targets, relations and landmarks is also estimated. As shown in Figure 5 this analysis reveals that spatial relations overall receive less visual attention than targets and landmarks. It is also interesting that landmarks receive slightly less visual attention than targets, perhaps because they can be predicted from the sequential language model. The visual attention for relations is not only weaker but also more dispersed which indicates

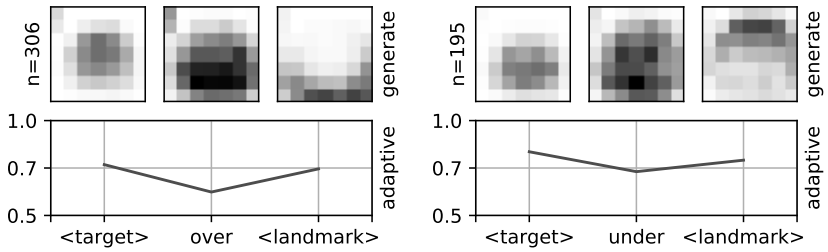


Figure 5: Average visual and language attention over a grid of 7×7 locations in images. Darker values indicate stronger visual attention and weaker language attention. The shades are relative to each other only within a box.

that it is not capturing geometric arrangements between target and landmark that we observe in spatial templates but is driven by some other factors, possibly related to the representations of targets and landmarks.

Visual features from pre-trained object classifiers are therefore useful for identifying objects but not relations between objects which are not directly identified by visual features. What happens in the same model when geometric features about the objects are explicitly added and a grounded language model is allowed to attend over them as well? What is the optimal representation of geometry between targets and landmarks? Ghanimifard and Dobnik (2019b) extend the previous model of adaptive attention to a model that can attend different representations of the scene: (a) visual features from bounding boxes of annotated objects, (b) ordered visual features of the target and landmark bounding boxes, (c) ordered features representing geometric relations between the target and landmark objects. The models are compared to two baseline models: (i) model without attention and (ii) a model with adaptive attention over the visual features of the entire scene. For training and evaluation again Visual Genome is used (Krishna et al., 2017). The descriptions generated from these annotations are therefore simple descriptions containing `<s> target relation landmark </s>`. A comparison of cross-entropy loss of different models indicates that top-down infused knowledge about objects and their geometric arrangement has a positive effect compared to using purely bottom-up knowledge. Overall, (b) identification of visual features of targets and landmarks leads to the largest improvement, greater than (c) geometric relations between targets and landmarks. When (b) and (c) are used together this leads to a further but small improvement over (b). Difference between relations are observed which suggests that not all features are equally relevant for all relations.

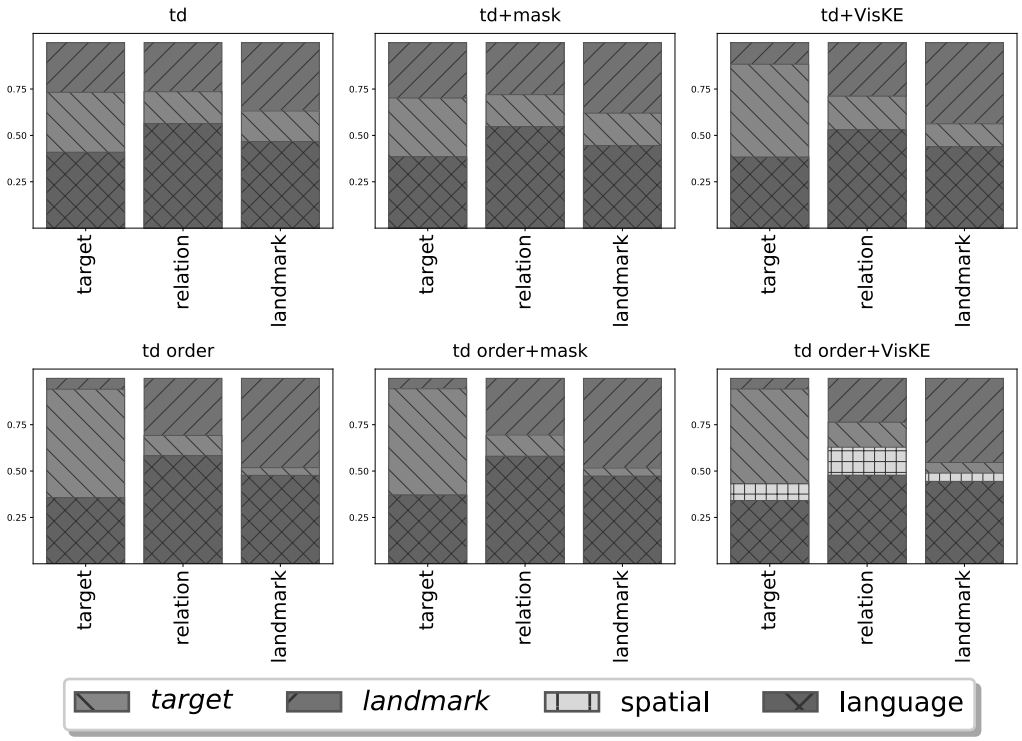


Figure 6: The average adaptive attention β on targets, relations and landmarks for different models. td: visual features within object bounding boxes, td order: ordered visual features of target and landmark objects, +mask: added masks of object bounding boxes as geometric features, +VisKE: added geometric relations between bounding boxes from VisKE (Sadeghi et al., 2015)

Figure 6 shows adaptive attention for individual parts of a description in respect of different features. The results indicate that only in the second case (td order), where the visual features of a target and a landmark are explicitly identified, geometric spatial features are used and that VisKE features (Sadeghi et al., 2015) which represent geometric properties and relations between target and landmark bounding boxes are more effective than masked representations of object bounding boxes. Identification of the visual features of targets and landmarks (td order) is better than only providing visual features of objects (td) as this family of models is able to attend better the target and landmark descriptions. Overall, the investigation shows that integration of top-down features has a positive effect on the performance of grounded language models but different features have different effect on different classes of relations. The results indicate that the question of feature representation and optimal information fusion is still very much open for the current and future research on grounded vision and language models.

6 Conclusion

The background to this article is the ongoing debate relating to the level of language "understanding" large language models, that encode a distributional analysis of text corpora, can claim. We have proposed that situated spatial language is a useful use-case for this debate to consider. This proposal is based on the observation that the variety of information types that spatial language draws upon make it a challenging research topic for computational systems. In particular, successfully understanding and generating situated spatial language requires the ability to blend geometric, perceptual, and functional/world knowledge. A key argument put forward in this article is that distributional semantics may provide an approach that is particularly suited to learning and encoding the functional/world knowledge information necessary to process spatial language. In support of this argument, in the later half of this article we have showcased a number of works that have demonstrated how distributional analysis (be it via an entropy based analysis of image captions, or the behaviour of language models) can reveal aspects of spatial semantics related to function. Much more work is necessary to understand precisely how distributional analysis can be best integrated with geometric and perceptual models to fully do justice to the richness of situated spatial semantics. However, we believe that recognising that distributional analysis provides basis to learning functional relationships between objects provides a useful direction for future work on spatial language.

Acknowledgements

The research of Kelleher was supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106_P2) and is co-funded under the European Regional Development Fund. The research of Dobnik was supported by a grant from the Swedish Research Council (VR project 2014–39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Lawrence W Barsalou et al. 1999. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(6):1137–1155.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735.

Patrick Blackburn and Johan Bos. 2005. *Representation and inference for natural language. A first course in computational semantics*. CSLI Publications.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Michael Brenner, Nick Hawes, John D Kelleher, and Jeremy L Wyatt. 2007. Mediating between qualitative and quantitative representations for task-orientated human-robot interaction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2072–2077.

Laura A Carlson-Radvansky and David E Irwin. 1993. Frames of reference in vision and language: Where is above? *Cognition*, 46(3):223–244.

Laura A Carlson-Radvansky and David E Irwin. 1994. Reference frame activation during spatial term assignment. *Journal of memory and language*, 33(5):646–671.

Laura A Carlson-Radvansky and Gordon D Logan. 1997. The influence of reference frame selection on spatial template construction. *Journal of memory and language*, 37(3):411–437.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#∗ vector: Probing sentence embeddings for linguistic properties. In *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2126–2136. Association for Computational Linguistics.

Fintan J Costello and John Kelleher. 2006. Spatial prepositions in context: The semantics of near in the presence of distractor objects. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*.

Kenny R Coventry and Simon C Garrod. 2004. *Saying, seeing and acting: The psychological semantics of spatial prepositions*. Psychology Press.

Kees van Deemter. 2010. *Not Exactly: In Praise of Vagueness*. Oxford University Press.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik and Amelie Åstbom. 2017. (Perceptual) grounding as interaction. In *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, pages 17–26, Saarbrücken, Germany.

Simon Dobnik and Mehdi Ghanimifard. 2020. Spatial descriptions on a functional-geometric spectrum: the location of objects. In *Spatial Cognition XII: Proceedings of the 12th International Conference, Spatial Cognition 2020, Riga, Latvia*, pages 219–234, Cham, Switzerland. Springer International Publishing.

Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Simon Dobnik and Erik de Graaf. 2017. KILLE: a framework for situated agents for learning language through interaction. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 162–171, Gothenburg, Sweden. Northern European Association for Language Technology (NEALT), Association for Computational Linguistics.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden.

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSci 2013 Production of referring expressions – bridging the gap*

between cognitive and computational approaches to reference at CogSci, pages 1–6, Berlin, Germany.

Simon Dobnik and John D. Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third V&L Net Workshop on Vision and Language at COLING*, pages 33–37, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.

Simon Dobnik, John D. Kelleher, and Christine Howes. 2020. Local alignment of frame of reference assignment in English and Swedish dialogue. In *Spatial Cognition XII: Proceedings of the 12th International Conference, Spatial Cognition 2020, Riga, Latvia*, pages 251–267, Cham, Switzerland. Springer International Publishing.

Simon Dobnik, John D. Kelleher, and Christos Koniaris. 2014. Priming and alignment of frame of reference in situated conversation. In *Proceedings of DialWatt – Semdial 2014: The 18th Workshop on the Semantics and Pragmatics of Dialogue*, pages 43–52, Edinburgh.

Simon Dobnik and Vera Silfversparre. 2021. The red cup on the left: Reference, coreference and attention in visual dialogue. In *Proceedings of PotsDial - Semdial 2021: The 25th Workshop on the Semantics and Pragmatics of Dialogue*, Proceedings (SemDial), pages 50–60, Potsdam, Germany.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

John R Firth. 1957. A synopsis of linguistic theory, 1930–1955. *Studies in linguistic analysis*, Special volume of the Philological Society:1–32.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*.

Klaus-Peter Gapp. 1994. Basic meanings of spatial relations: Computation and evaluation in 3d space. In *AAAI*, pages 1393–1398. AAAI Press/The MIT Press.

Gerald Gazdar. 1996. Paradigm merger in natural language processing. In Ian Wand and Robin Milner, editors, *Computing tomorrow: future research directions in computer science*, pages 88–109. Cambridge University Press.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In *Proceedings of IWCS 2017: 12th International Conference on Computational Semantics*, pages 1–12, Montpellier, France. Association for Computational Linguistics.

Mehdi Ghanimifard and Simon Dobnik. 2018. Knowing when to look for what and where: Evaluating generation of spatial descriptions with adaptive attention. In *Computer Vision – ECCV 2018 Workshops. ECCV 2018*, volume 11132 of *Lecture Notes in Computer Science (LNCS)*, pages 1–9, Proceedings of the Workshop on Shortcomings in Vision and Language (SiVL), ECCV 2018, Munich, Germany. Springer, Cham.

Mehdi Ghanimifard and Simon Dobnik. 2019a. What a neural language model tells us about spatial relations. In *Proceedings of the Combined*

Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP), pages 71–81, Minneapolis, Minnesota, USA. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Association for Computational Linguistics.

Mehdi Ghanimifard and Simon Dobnik. 2019b. What goes into a word: generating image descriptions with top-down spatial knowledge. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG-2019)*, pages 1–15, Tokyo, Japan. Association for Computational Linguistics.

Peter Gorniak and Deb Roy. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

Michael Grubinger, Paul D. Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR benchmark: A new evaluation resource for visual information systems. In *Proceedings of OntoImage 2006: Workshop on language resources for content-based mage retrieval during LREC 2006*, Genoa, Italy. European Language Resources Association.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Nick Hawes, Matthew Klenk, Kate Lockwood, Graham Horn, and John Kelleher. 2012. Towards a cognitive system that can recognize spatial regions based on context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, 32:11137–11147.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge.

Geoffrey E. Hinton, James L. McClelland, and David E. Rumelhart. 1986. Distributed representations. In David E. Rumelhart, James L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing*, volume 1, chapter 3, pages 77–109. MIT Press.

Nikolai Ilinykh and Simon Dobnik. 2021a. How vision affects language: Comparing masked self-attention in uni-modal and multi-modal transformer. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR) at IWCS 2021*, pages 45–55, Groningen, Netherlands (Online). Association for Computational Linguistics.

Nikolai Ilinykh and Simon Dobnik. 2021b. What does a language-and-vision transformer see: The impact of semantic information on visual representations. *Frontiers in Artificial Intelligence: Identifying, Analyzing, and Overcoming Challenges in Vision and Language Research*, 4(767971):182–203.

Nikolai Ilinykh and Simon Dobnik. 2022. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, Dublin, Ireland. Association for Computational Linguistics.

Ray Jackendoff and Barbara Landau. 1991. Spatial language and spatial cognition. In *Bridges between psychology and linguistics*, pages 157–182.

Psychology Press.
Ray S Jackendoff. 1985. *Semantics and cognition*, volume 8. MIT press.
Magdalena Kacmajor and John D Kelleher. 2020. Capturing and measuring thematic relatedness. *Language Resources and Evaluation*, 54(3):645–682.
Magdalena Kacmajor, John D Kelleher, Filip Klubicka, and Alfredo Maldonado. 2020. Semantic relatedness and taxonomic word embeddings. *arXiv preprint arXiv:2002.06235*.
Jussi Karlgren and Pentti Kanerva. 2021. Semantics in high-dimensional space. *Frontiers in Artificial Intelligence*, 4:123.
John Kelleher and Fintan J Costello. 2005. Cognitive representations of projective prepositions. In *Proceedings of the Second ACL-Sigsem Workshop of The Linguistic Dimensions of Prepositions and their Use in Computational Linguistic Formalisms and Applications*.
John Kelleher and Josef van Genabith. 2004. Visual salience and reference resolution in simulated 3-d environments. *Artificial Intelligence Review*, 21(3):253–267.
John Kelleher and Josef van Genabith. 2006. A computational model of the referential semantics of projective prepositions. In *Syntax and Semantics of Prepositions*, pages 211–228. Springer.
John Kelleher and Geert-Jan M Kruijff. 2005a. A context-dependent algorithm for generating locative expressions in physically situated environments. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.
John Kelleher and Geert-Jan M Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*, pages 1041–1048.
John Kelleher, Geert-Jan M Kruijff, and Fintan J Costello. 2006. Proximity in context: an empirically grounded computational model of proximity for processing topological spatial expressions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 745–752.
John Kelleher, Colm Sloan, and Brian Namee. 2009. An investigation into the semantics of English topological prepositions. *Cognitive Processing*, 2(10):233–236.
John D. Kelleher. 2003. *A perceptually based computational framework for the interpretation of spatial language*. Ph.D. thesis, Dublin City University.
John D. Kelleher. 2006. Attention driven reference resolution in multimodal contexts. *Artificial Intelligence Review*, 25(1):21–35.
John D Kelleher. 2019. *Deep learning*. MIT press.
John D Kelleher and Fintan J Costello. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306.
John D. Kelleher, Fintan J. Costello, and Josef van Genabith. 2005. Dynamically structuring updating and interrelating representations of visual and linguistic discourse. *Artificial Intelligence*, 167(1):62–102.
John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *Conference on Logic and Machine Learning in Natural Language (LaML*

2017), volume 1 of *CLASP Papers in Computational Linguistics*, pages 41–52.

John D Kelleher and Simon Dobnik. 2019. Referring to the recently seen: reference and perceptual memory in situated dialogue. In *Extended papers from the Conference on Dialog and Perception*, volume 2 of *CLASP Papers in Computational Linguistics*, pages 41–50.

John D Kelleher and Geert-Jan M Kruijff. 2005b. A context-dependent model of proximity in physically situated environments. In *Proceedings of the 2nd ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational*.

John D Kelleher, Robert J Ross, Colm Sloan, and Brian Mac Namee. 2011. The effect of occlusion on the semantics of projective spatial terms: a case study in grounding language in perception. *Cognitive Processing*, 12(1):95–108.

Filip Klubička, Alfredo Maldonado, Abhijit Mahalunkar, and John Kelleher. 2019. Synthetic, yet natural: Properties of wordnet random walk corpora and the impact of rare words on embedding performance. In *Proceedings of the 10th Global Wordnet Conference*, pages 140–150.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Geert-Jan M Kruijff, John D Kelleher, Gregor Berginc, and Aleš Leonardis. 2006a. Structural descriptions in human-assisted robot visual learning. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 343–344.

Geert-Jan M Kruijff, John D Kelleher, and Nick Hawes. 2006b. Information fusion for visual reference resolution in dynamic situated dialogue. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 117–128. Springer.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. The University of Chicago Press, Chicago and London.

Barbara Landau. 1996. Multiple geometric representations of objects in languages and language learners. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and space*, A Bradford Book, chapter 8, pages 317–363. The MIT Press, Cambridge, Massachusetts and London, England.

Barbara Landau and Ray Jackendoff. 1993. "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238.

Stephen C Levinson. 1996. Frames of reference and Molyneux's question: Crosslinguistic evidence. *Language and space*, 109:169.

Stephen C Levinson. 2003. *Space in language and cognition: explorations in cognitive diversity*, volume 5. Cambridge University Press, Cambridge.

Xiaoou Li, Laura A Carlson, Weimin Mou, Mark R Williams, and Jared E Miller. 2011. Describing spatial locations from perception and memory: The influence of intrinsic axes on reference object selection. *Journal of Memory and Language*, 65(2):222–236.

Annika Lindh, Robert Ross, and John Kelleher. 2020. Language-driven region pointer advancement for controllable image captioning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1922–1935.

Annika Lindh, Robert J Ross, Abhijit Mahalunkar, Giancarlo Salton, and John D Kelleher. 2018. Generating diverse and meaningful captions. In *International Conference on Artificial Neural Networks*, pages 176–187. Springer.

Gordon D Logan. 1994. Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20(5):1015.

Gordon D Logan. 1995. Linguistic and conceptual control of visual spatial attention. *Cognitive psychology*, 28(2):103–174.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208.

Didier Maillat. 2003. *The semantics and pragmatics of directionals: a case study in English and French*. Ph.D. thesis, University of Oxford: Committee for Comparative Philology and General Linguistics, Oxford, United Kingdom.

Alfredo Maldonado, Filip Klubička, and John Kelleher. 2019. Size matters: The impact of training size in taxonomically-enriched word embeddings. *Open Computer Science*, 9(1):252–267.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

George A. Miller and Philip N. Johnson-Laird. 1976. *Language and perception*. Cambridge University Press, Cambridge.

Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. 1983. Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417.

Richard Montague. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven. Ed. and with an introduction by Richmond H. Thomason.

Vasudevan Nedumpozhimana and John Kelleher. 2021. Finding BERT's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62.

Vasudevan Nedumpozhimana, Filip Klubicka, and John D. Kelleher. 2022.

Shapley idioms: Analysing BERT sentence embeddings for general idiom token identification. *Frontiers in Artificial Intelligence*, 5.

Alva Noë. 2004. *Action in perception*. The MIT Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on creating speech and language data with Amazon's Mechanical Turk*, Los Angeles, CA. North American Chapter of the Association for Computational Linguistics (NAACL).

Terry Regier and Laura A. Carlson. 2001. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464.

Magnus Sahlgren and Fredrik Carlsson. 2021. The singleton fallacy: Why current critiques of language models miss the point. *Frontiers in Artificial Intelligence*, 4:131.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204.

Giancarlo Salton, Robert Ross, and John Kelleher. 2017. Attentive language models. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 441–450.

Holger Schultheis and Laura A Carlson. 2017. Mechanisms of reference frame selection in spatial term use: computational and empirical studies. *Cognitive science*, 41(2):276–325.

Holger Schultheis and Laura A. Carlson. 2018. Inter-process relations in spatial language: Feedback and graded compatibility. *Cognition*, 176:140–158.

Niels Schütte, Brian Mac Namee, and John Kelleher. 2017. Robot perception errors and human resolution strategies in situated human–robot dialogue.

Advanced Robotics, 31(5):243–257.

Hinrich Schütze. 1993. Word space. In *Advances in neural information processing systems*, pages 895–902.

Leonard Talmy. 1983. How language structures space. In *Spatial orientation*, pages 225–282. Springer.

Leonard Talmy. 2000. *Toward a cognitive semantics: concept structuring systems*, volume 1 and 2. MIT Press, Cambridge, Massachusetts.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Anne Treisman and Stephen Gormican. 1988. Feature analysis in early vision: evidence from search asymmetries. *Psychological review*, 95(1):15.

Shimon Ullman. 1980. Against direct perception. *Behavioral and Brain Sciences*, 3(3):373–381.

Shimon Ullman. 1984. Visual routines. *Cognition*, 18(1-3):97–159.

Claude Vandeloise. 1991. *Spatial prepositions: A case study from French*. University of Chicago Press.

Dominic Widdows. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 197–204. Association for Computational Linguistics.

Terry Winograd. 1976. *Understanding Natural Language*. Edinburgh University Press.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10685–10694.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699.