(Perceptual) grounding as interaction

Simon Dobnik[*†] and **Amelie Åstbom**[*]
[*]Department of Philosophy, Linguistics and Theory of Science
[†]Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
simon.dobnik@gu.se, amelie.astbom@hotmail.com

Abstract

We examine how changing perceptual contexts affects grounding of words, in particular spatial descriptions, in perceptual features and argue that grounding is interactive. We discuss two effects of perceptual context. Grounding of spatial descriptions may be affected by the richness of the perceptual context which allows us to build more complex representations of scenes. Secondly, perceptual grounding is dependent on the task (and associated attention) which affects the preference of features. The second property connects perceptual grounding closely to linguistic grounding in dialogue. We argue that dynamic perceptual grounding has implications for the words-as-classifiers approach to semantics.

1 Introduction

Humans interact with each other and language is a central part of their interaction (Clark, 1996). The properties of their interaction define the semantics of words. There is a significant body of research that shows how semantics of words is coordinated and integrated in the common ground of conversational partners (Clark and Wilkes-Gibbs, 1986). The focus of these investigations is the linguistic interaction between conversational partners, the conversational strategies that they employ while observing and discussing a shared perceptual scene. However, this is only one part of the interaction that takes place in this scenario. Both conversational partners also interact with their environment through perception while constructing their representation of space. An important question is how do conversational partners know what properties of the environment are relevant when they generate or hear a description such as "the ball is over the basket". The properties of the perceptual scene, the features that agents individually and later through linguistic coordination consider as salient have an effect on the meaning assigned to words used in that context. In lexical semantics the idea of dynamic interpretation of words in contexts and defining procedures for generating semantic representations for words on the basis of particular contexts has been captured in the notion of the generative lexicon (Pustejovsky, 1995). The question of feature salience and selection has been mainly explored in the literature on generating referring expressions (GRE) (Dale and Reiter, 1995; Deemter, 2016). However, these models typically assume the features (and therefore semantics of referring expressions) are constant over all perceptual scenes. Extending this work we argue (based on the finding of our experimental results and inline with the notion of the generative lexicon) that feature selection is dynamic, dependent on (i) the feature richness of the perceptual scene which allows us to construct different representations of the scene and (ii) the task that an agent is engaged with which affects the salience of features. This poses a challenge to the view of grounding as classifiers (Harnad, 1990; Roy, 2005; Dobnik, 2009; Larsson, 2013; Schlangen et al., 2016) as these typically consider a fixed set of features that ground the semantics of expressions.

2 Spatial descriptions

We work in the domain of spatial descriptions such as "over", "above" and "left" and the composed phrases containing them such as "the ball is over and to the right of the basket". Spatial descriptions are a good domain because they are relatively complex phrases which include both references to objects and relations between objects. Studies in

spatial language (Herskovits, 1986; Talmy, 2000) show that their semantics are dependent on several contextual sources of information which can be briefly summarised as: (i) geometric arrangements of objects in the scene; (ii) properties of objects and properties of their interaction which can be modelled as conceptualisations in terms of geometric shapes and dynamic-kinematic routines over them; and (iii) the perspective from which the scene is described which determines the orientation of the geometric coordinate frame. Here, we focus on the first two and which we describe below.

The geometric representation of spatial descriptions can be represented by spatial templates which were introduced in (Logan and Sadler, 1996). Spatial templates denote degrees of acceptability of a particular description over two dimensional space (as such they are 3 dimensional graphs). In (Logan and Sadler, 1996) they are induced experimentally by designing a grid of 7×7 cells which is invisible to participants. The landmark object is always placed in the centre cell while the target object is placed in all other locations, one at a time. The locations encode three different degrees of distance away from the landmark in each direction. Participants are presented with pictures of such visual scenes and a particular description such as "The circle is above the box". Their task is to rate on a scale to what degree a given description matches the scene. The images are presented in a random order. Aggregating the average acceptability score per individual locations allows us to define regions of acceptability or grounding of that spatial description in space.

The effect of the properties of objects and their interaction has been studied in (Coventry et al., 2001; Coventry et al., 2005). They compare the spatial descriptions *over/under* and *above/below*. In (Logan and Sadler, 1996) objects are represented as abstract shapes and therefore their spatial templates look very similar since only the geometric dimension is taken into account. In the first experiment (Coventry et al., 2001) use scenes with functionally related objects (a man holding an umbrella) in different geometric configurations and alternate whether the functional relationship is fulfilled (with and without rain). The results show that *above/below* are more influenced by geometry while *over/under* are more influenced by function (the umbrella providing protection from the rain). In the second experiment, they introduce functionally inappropriate objects (a man holding a suitcase instead of an umbrella). The results are the same as for the first experiment but it is also the case that functionally appropriate scenes are rated higher than inappropriate ones but this does not interact with any of the main variables of interest. In the third experiment they show that in the scenes where the intrinsic and extrinsic reference frames do not coincide, this negatively affects the ratings for *above/below* while *over/under* are acceptable but only in those cases where the functional relation between the objects is fulfilled. (Hörberg, 2008) shows similar results for Swedish *över/under* and *ovanför/nedanför* with the exception that *under* and *nedanför* are not influenced by function to a different degree. This suggests that there are some cross-linguistic differences.

Both (Coventry et al., 2001) and (Hörberg, 2008) use sets of images representing functionally interacting or non-interacting objects with some variation of their location. (Hörberg, 2008) also shows that function influences acceptability regions depending on the properties of the interacting objects and compares them with the predictions of the Attentional Vector Sum Model (Regier and Carlson, 2001). In this paper we undertake a similar investigation by examining how expressions are grounded in spatial templates of (Logan and Sadler, 1996). A similar investigation of the effects of the context on the grounding of spatial descriptions in spatial templates has been performed in (Kelleher et al., 2006) but to study the effect of distractor objects. In particular we want to answer the following questions:

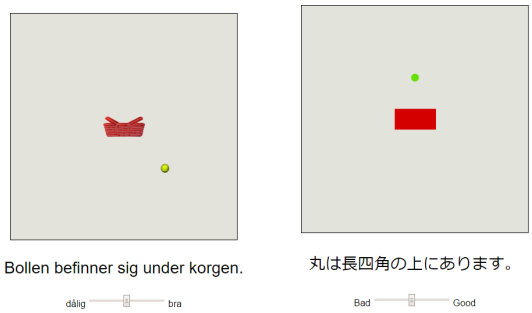
1. Do physical properties of the environment, the representation of objects related by a spatial description, have an effect on its semantic interpretation measured in terms of its grounding in a spatial template? If such an effect is shown, then the semantics of spatial descriptions, their grounding in spatial templates, is not static but is being constantly defined by the perceptual context.
2. We expect grounding (the semantics of spatial expressions) to be also affected by their distributional properties (Turney et al., 2010), how they are used in a particular language in general. It follows that there will be differences in grounding of words belonging to

different languages, in our case Swedish and Japanese.

3. Words compose to form phrases. Is grounding compositional in the same way as predicted by formal compositional semantics (Blackburn and Bos, 2005)? If this is so, then functional composition of words should be reflected in a (predictable) functional composition at the level of spatial templates. Can the grounding of complex descriptions be predicted from the grounding of simple descriptions? Or is composition also dynamic?

3 Experiment

Two sets of images of perceptual situations were produced. In the first set of images the target and landmark objects are geometric shapes (a rectangle and a circle) while in the second set they are images of objects (a basket and a ball). The ball and a basket can interact in several ways. For example, the basket can be seen as a container to capture the ball or to provide protection/coverage for the ball. Geometric shapes are simpler representations than drawings of objects which means that they will allow for different conceptualisation of the spatial relation between the objects (also bringing in different functional knowledge)[1] and therefore we expect that they will have different effect on the grounding of the spatial description that they are relating.



(a) (b)

Figure 1: The experiment task for (a) geometric and (b) functional context. Descriptions: (a) The ball is under the basket. (b) The circle is over the rectangle.

To investigate the effects of different language models we compare the grounding of the corresponding expressions to *over* and *under*, two spatial descriptions that have been shown to be sensitive to function (Coventry et al., 2001; Hörberg, 2008), in Swedish and Japanese. Swedish makes a similar distinction between function-sensitive (*över/under*) and geometry-sensitive (*ovanför/nedanför*) pairs as English whereas in Japanese there is no such distinction (上/下: *ue/shita*). English/Swedish descriptions will therefore have different distributional properties from Japanese.

To investigate the compositionality of grounding of composed spatial descriptions we compare artificially composed spatial templates by some known function with a spatial template of a "naturally" composed description obtained experimentally in the same perceptual context. In particular, we compare two different compositions of Swedish "över" + "vänster" (over + left) with "över och till vänster" (over and to the left).

3.1 Task

Three experiments were performed. In Experiment 1 we collect judgements for Swedish *över/under* in geometrical and functional contexts. In Experiment 2 we collect judgements for Japanese (上/下: *ue/shita*) in geometric and functional contexts and in Experiment 3 we collect judgements for Swedish "naturally" composed descriptions in the functional context. Spatial descriptions are embedded within a sentence also containing descriptions of the related objects.

We use an online tool for collection of linguistic data called Semant-O-Matic that we developed ourselves and has been used in several other tasks.[2] Its benefit in comparison to other crowdsourcing tools such as Amazon Mechanical Turk (AMT) is that it allows us a better control of participants, speakers of Swedish and Japanese, by distribution of sign-up links. Random participation is prevented by requiring each participant to provide a valid e-mail address. The requirement to be a native speaker of a language was strengthened by having instructions in Swedish but this was not the case for Japanese where instructions were in English. After signing up, each participant received an email with experimental instructions and a personal link to the experiment. The tool is therefore a convenient compromise between a lab experiment and an open crowd-sourcing scenario.

Participants were randomly assigned either to

[1] Perceptual and encyclopedic world-knowledge features of objects are closely linked together.

[2] http://www.dobnik.net/simon/semant-o-matic/

the geometric or functional perceptual contexts (Experiment 1 and 2). For Experiment 3, participants who have already taken part in Experiment 1 were re-invited. For Experiment 1 and 2 we choose a between-subject design of the experiment for each language rather than a within-subject design because the latter would explicitly introduce a distinction between these two contexts. This way, we kept it open for participants to decide how to interpret each spatial context. Preserving the perceptual contexts is also important if our task is to capture an entire spatial template for that context which can be applied in description generation and interpretation. Figure 1 shows an example of the task in both perceptual contexts and for both languages. For each presentation, a participant's task was to move the slider below the image between the two extremes (bad and good) in order to indicate how appropriate the description is for that scene. The slider translated to an underlying scale ranging from 0 to 100 but this was not visible to the participant.[3] The images with different location of the target object relative to the landmark were presented in a random order. In each Experiment 1, 2 and 3 we were testing two descriptions which means that they contained a total of $48 \times 2 = 96$ presentations.

3.2 Participants

Experiment 1 was completed by 29 participants, 13 of whom were assigned the geometric context and 16 of whom were assigned the functional context. If a participant did not complete all 48 judgements for a spatial template, their score was replaced by the mean score of other participants per that context and location. The number of responses for the functional context ranged between 13 to 16 and the number of responses for the geometric context ranged between 12 and 13. All participants completed the experiment but there were occasional missing values. Experiment 2 was attempted by 8 participants with 4 participants per each context. The number of responses for the functional context ranged between 3 and 4 (complete responses with an occasional missing value) and the number of responses for the geometric context ranged from 2 to 4 (2 participants only partially completed the experiment). Experiment 3 was attempted by 12 participants of whom 1 only partially completed it.

[3] In this respect our scenario differs from (Logan and Sadler, 1996) who use a scale of numbers from 1 to 9.

4 Data and analysis

As stated earlier, for each spatial description and for each context in which it was used we calculate a mean acceptability rating per each of the 48 locations. The means form a spatial template. To quantitatively evaluate the difference between individual spatial templates we use a Wilcoxon signed-rank test and Pearson's correlation coefficient (r) over these 48 means.

4.1 The effect of perceptual context

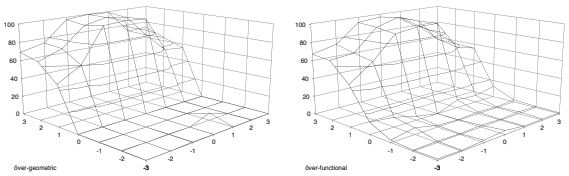


Figure 2: The spatial template for Swedish "över" in geometric and functional contexts.

Figure 2 shows a spatial template for Swedish **"över"** in both contexts. Surprisingly, they appear very similar. A Wilcoxon signed-rank test found no significant difference between *över-geometric* and *över-functional* ($V = 481, p = 0.383$). The mean scores per location are also highly correlated ($r(46) = 0.995, p < 0.001$) which is also shown in plot in Figure 3. Our hypothesis that there will be an effect of the perceptual context on the grounding of a spatial description is therefore not supported in this case.

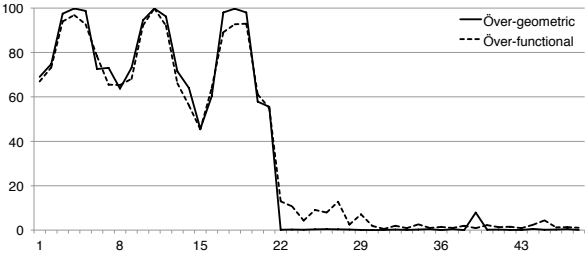


Figure 3: Variation of mean acceptability scores for "över". Each cycle of 7 represents one row in a spatial template.

Figure 4 shows a spatial template for Swedish **"under"** in geometric and functional contexts. A Wilcoxon signed-rank test found no significant difference between *under-geometric* versus *under-functional* ($V = 445, p = 0.145$). The data is

also highly correlated ($r(46) = 0.969, p < 0.001$). Again, the hypothesis that there is an effect of the perceptual context on the grounding of a spatial description is not supported.

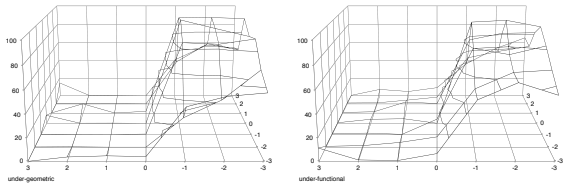


Figure 4: The spatial template for Swedish "under" in geometric and functional contexts rotated by 90° anticlockwise.

Let us now turn to Japanese. As mentioned earlier, Japanese does not distinguish between *over/above* and *under/below* as English and Swedish do. However, 上 "ue" (over/above) and 下 "shita" (under/below) could still show different effects on grounding in functional and geometric contexts. Figure 5 shows the spatial templates for 上 **"ue"**. A visual observation reveals that in the geometric context the acceptability rating decrease more rapidly away from the centre of the scene and that more unexpected (yet low) acceptability ratings are found in the geometric but not functional context ($y < 0$). In this case a Wilcoxon signed-rank test found a significant difference between *ue-geometric* and *ue-functional* ($V = 867, p < 0.001$). The data is highly correlated ($r(46) = 0.961, p < 0.001$).

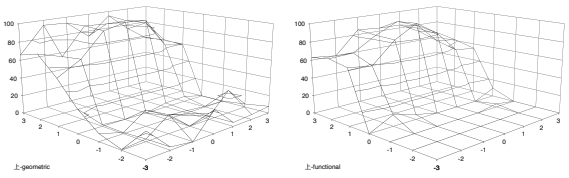


Figure 5: The spatial template for Japanese 上 "ue" in geometric and functional contexts.

Figure 6 shows the spatial templates for Japanese 下 **"shita"**. A visual observation reveals that the acceptability scores for the functional context are overall lower than the scores for the geometric context. The scores in the functional context decrease more steeply from the centre position (not visible in this graph). Similarly to the previous comparison involving 上 "ue", a Wilcoxon signed-rank test found a significant difference between *shita-geometric* and *shita-functional* ($V = 785, p < 0.001$). The data is also highly correlated ($r(46) = 0.923, p < 0.001$).

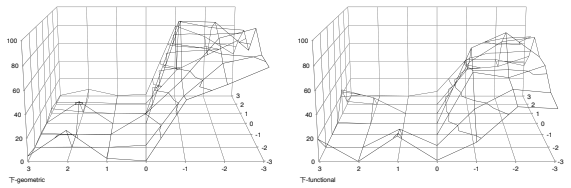


Figure 6: The spatial template for Japanese 下 "shita" in geometric and functional contexts.

Overall, the results presented in this section show that there is no effect of the perceptual context on the grounding of "över" and "under" in Swedish, while there is an effect on the grounding of 上 "ue" and 下 "shita" in Japanese.

4.2 The effect of the language model

In this section we examine grounding of parallel descriptions across different languages. Let us first consider grounding of descriptions in the geometric context. Figure 7 shows spatial templates for "över" and 上 "ue" in the geometric context. A Wilcoxon signed-rank test found a significant difference between *över-geometric* and *ue-geometric* ($V = 360.5, p = 0.02$). The data is also highly correlated ($r(46) = 0.970, p < 0.001$).

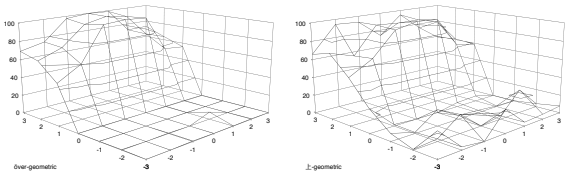


Figure 7: The spatial template for Swedish "över" and Japanese 上 "ue" in the geometric context.

Figure 8 shows the spatial templates for "under" and 下 "shita" in the geometric context. A Wilcoxon signed-rank test found no significant difference between *under-geometric* and *shita-geometric* ($V = 436, p = 0.120$). The data is also highly correlated ($r(46) = 0.944, p < 0.001$).

Let us now turn to the grounding of parallel spatial descriptions across different languages in the functional context. Figure 9 shows the spatial templates for "över" and 上 "ue" in the functional context. A visual comparison reveals that 上 "ue" is more sensitive to proximity to the centre or the x-axis. A Wilcoxon signed-rank test found a significant difference between *over-functional* and *ue-*

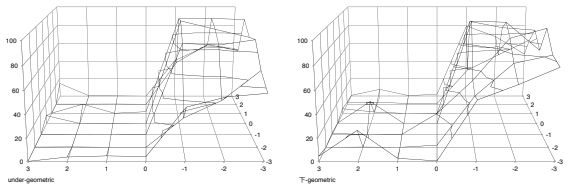


Figure 8: The spatial template for Swedish "under" and Japanese 下 "shita" in the geometric context.

functional ($V = 997, p < 0.001$). The data are also highly correlated ($r(46) = 0.991, p < 0.001$).

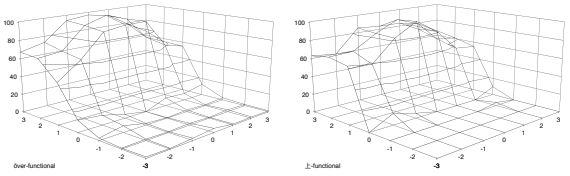


Figure 9: The spatial template for Swedish "över" and Japanese 上 "ue" in the functional context.

Finally, Figure 10 shows the spatial templates for "under" and 下 "shita" in the functional context. The graphs show that 下 "shita" has overall lower acceptability scores that "under" and that the latter are more varied. A Wilcoxon signed-rank test found a significant difference between *under-functional* and *shita-functional* ($V = 971, p < 0.001$). The data are also highly correlated ($r(46) = 0.947, p < 0.001$).

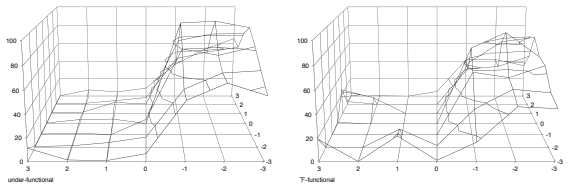


Figure 10: The spatial template for Swedish "under" and Japanese 下 "shita" in the functional context.

4.3 The effect of word composition

In this section we explore whether spatial templates of complex phrases or composite spatial descriptions can be predicted from the spatial templates of the individual words that are a part of a composite description. Can the grounding of phrases be seen as a composition of functions in model theoretic semantics or is it interactive depending both on the grounding and distributional properties of individual words? There has been significant focus on the question of semantic composition in computational semantics but the investigations focus on the composition of vector spaces (thus distributions of words in their contexts) rather than composition of grounded representations of words in the physical world (Mitchell and Lapata, 2010; Clark, 2015). Here, composition can be achieved by some mathematical operation on distributional tensors (higher-order vectors representing distributional contexts of words), typically multiplication.

We investigate the semantic interaction of composed words in phrases in terms of their grounding by comparing the grounding of artificially composed spatial templates of individual words with a "naturally" grounded spatial template of a composite description. In particular we examine the Swedish description "över och till vänster om" (over and to the left of) in the functional context. We already obtained the spatial template for "över" in Experiment 1 and hence in Experiment 3 we collect spatial templates for "till vänster om" (to the left of) and "over och till vänster om". Figure 11 shows the spatial templates of the individual words.

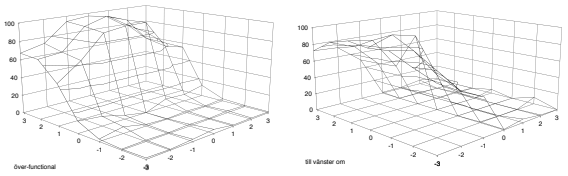


Figure 11: The spatial template for Swedish "över" and "till vänster om" in the functional context.

For artificial composition we test two compositional functions: *arithmetic mean* ($\frac{a+b}{2}$) and *geometric mean* ($\sqrt[2]{a \times b}$). Since both functions are types of mean they ensure that the composed values are within the same range as the values before the composition which means that the scores can be directly compared.

Figure 12 shows a comparison of both artificially grounded compositions with the natural grounding of the composed phrase. As evidenced by the later, the highest acceptability ratings concentrate in the first quadrant where $x < 0, y > 0$. It follows from the visual observation that geometric mean is a better compositional function for spatial templates than arithmetic mean as the latter also predicts undesirable acceptable

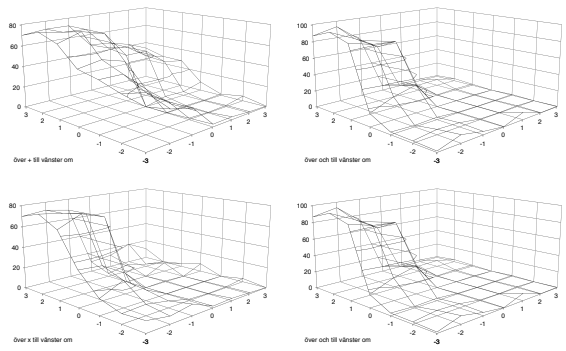


Figure 12: The spatial template for Swedish "över" $+/\times$ "till vänster om" and "över och till vänster om" in the functional context.

regions in the quadrants 2 ($x > 0, y > 0$) and 3 ($x < 0, y < 0$). A Wilcoxon signed-rank test found a significant difference between "över" + "till" "vänster" (arithmetic) versus "över och till vänster" (natural) ($V = 185.5, p < 0.001$). The data are highly correlated but r is considerably lower than in the previous investigations ($r(46) = 0.781, p < 0.001$). In contrast, a Wilcoxon signed-rank test found no significant difference between "över" $\times$ "till vänster" (geometric) versus "över och till vänster" (natural) ($V = 562, p = 0.794$). These data are also highly correlated ($r(46) = 0.959, p < 0.001$). Hence, it follows that geometric mean as a compositional function approximates very well natural composition. (Gapp, 1994) discusses (but not experimentally evaluates) five compositional functions for grounding spatial templates and concludes that a *scaled minimum* of applicability scores preserves all the required properties of spatial templates under composition: $DA_{Rel_{cp}} := S(Min(DA_{Rel1}, DA_{Rel2})) \times Min(DA_{Rel1}, DA_{Rel2})$ where S is some contextually defined scaling factor. The first part of the equation ensures that S has a different effect on acceptability scores of different sizes. This compositional function is similar to geometric mean that we use. However, the latter is simpler and always ensures the scaling of the predicted acceptability score within the range of the original values.

Returning to the question of interaction of grounded semantics of spatial descriptions in composition, the findings suggest that this might be fixed as it can be predicted well by a simple mathematical function.

Description	p	Sig	r
Perceptual context: geometric vs functional			
över	0.383	ns	0.995
under	0.145	ns	0.969
上 ue	< 0.001	***	0.961
下 shita	< 0.001	***	0.923
Language: Swedish vs Japanese			
geo: över - 上 ue	0.02	*	0.970
geo: under - 下 shita	0.120	ns	0.944
func: över - 上 ue	< 0.001	***	0.991
func: under - 下 shita	< 0.001	***	0.947
Composition: artificial vs natural			
+: över och till vänster	< 0.001	***	0.781
$\times$: över och till vänster	0.794	ns	0.959

Table 1: Summary of comparisons

5 Discussion

Table 1 summarises the results of all comparisons. Let us first turn to our first question: do the properties of the perceptual context, the complexity of objects related by spatial relations *over* and *under* have an effect on the grounding of spatial templates. The results indicate that the perceptual context had influence on the grounding of words in Japanese but not in Swedish. The data from Japanese therefore confirms the previous findings for English and Swedish. However, the effect of the context on the spatial templates for Japanese should be taken with caution as the acceptability scores were collected from fewer participants and therefore the differences could be because of overfitting. On the other hand, our results for Swedish are surprising, because we know from (Hörberg, 2008) that "över" and "under" show sensitivity to functional relations between objects. However, there is an important difference in the way their (and Coventry et al.'s) and our tasks were structured, in particular the way stimuli were presented to participants. Participants in their study were exposed to a series of images that in terms of function could be classified to one of the following three categories: functional interaction, no-functional interaction and no-need for functional interaction. Therefore, the presence or absence of a strong functional interaction between objects was made a salient feature in their task. On the other hand, in our scenario, participants always provided judgements within one perceptual context and it was up to them to decide whether to take the functional interaction between the objects as a salient property of the context for the interpretation (while estimating a belief that this was the intention of the speaker of the utterance). This

means that perceptual grounding is dynamic and is constructed on the fly upon the evaluation of the scene and the linguistic discourse. A further support for this claim comes from the observation from one of our participants who interpreted the non-functional scene (involving abstract objects) as a functional scene, since in their view it resembled the game of Pong. It is also important to emphasise the relation of our findings to (Logan and Sadler, 1996). There the stimuli lacked functional dimension altogether as the data only contained objects of geometric shapes and for that reason only geometric dimension of the grounding could be taken into account. In our stimulus, the participants had a choice between the two but they appeared to have taken bias towards the geometric context while taking into account the functional context only weakly as the object function was not a salient feature of the task.

Our second question was whether we would expect a different behaviour in grounding of words belonging to different languages on the grounds of their distributional properties or their use in that language. Our findings indicate that there is a stronger difference between the Swedish and the Japanese descriptions in the functional context than in the geometric context. Coupled with the previous observation that the perceptual context had an effect on grounding of words in Japanese but not Swedish it appears that Japanese words are more adaptable to different contexts. Note that Japanese lacks a lexical distinction between functional/geometric pairs present in English and Swedish ("over"/"above" and "under"/-"below"). Therefore, Japanese 上 "ue" and 下 "shita" are used over a greater variety of situations than Swedish "över" and "under" (their grounding is more adaptable to contexts) while "över" and "under" are competing with "övanför" and "nedanför". The presence of a lexicalised sensitivity to object function in "över" may therefore make the grounding of "over" more stable or conservative across contexts. The contribution of word distributions in a language model is an interesting and open research question which we hope to address in the future.

Finally, our third question examines whether grounding is compositional in the same way as words are believed to be compositional in a language model. Our results indicate that composed grounding in a particular perceptual context can be predicted by a simple compositional function. This is important in respect to the previous findings that grounding of words or concepts is dynamic, depending on the context. If grounding of composed words were not predictable and also dynamic then it were far more difficult to interpret (and learn meanings of) composed phrases. Composition is therefore a property of the mechanics of language and not the lexicon. This conclusion is in line with the findings of (Kirby et al., 2008) on computational modelling of multiple generations of agents who show that compositionality of language emerges from language through repeated transmissions over generations through the learning bottle-neck: having learned compositional rules an agent can infer the language as a whole. Crucially, this requires compositionality to be constant across lexical variability. In order to confirm our hypothesis, we would have to investigate the grounded composition of words under different contexts and different pairs of lexical items, not just spatial relations. In a separate line of work (Ghanimifard and Dobnik, 2017) we carried out an experiment with machine learning of spatial descriptions grounded in spatial templates where the system is able to ground successfully "decomposed" descriptions while having learned only from their composed representations. This provides a further support for our current claim.

6 Conclusions and future work

The preceding discussion shows that perceptual grounding is dynamic and interactive. First, expressions may be grounded differently based on the number of available perceptual features in the current context. Secondly, the presence of a feature in a context is not always enough for that feature to be used in grounding of a description. There is a further selection of relevant and irrelevant features which is related to the task the conversational participants are performing. This way perceptual grounding can be seen as a dynamic negotiation of conversational participants with the environment. Of course, participants also negotiate through dialogue with other participants but that has to do with lexical choice which provides bias for perceptual grounding. The two interactive processes are therefore tightly connected.

Dynamic perceptual grounding has implications for building situated conversational agents. Most systems assume that agents use the same ground-

ing models or classifiers (although these may be incrementally learnable) over a variety of situations and even tasks. What the findings here suggest is that an agent would require a mechanism of attention that monitors perceptual and dialogue conversations and predicts a focus on certain features of both contexts that it can explore in grounding (Dobnik and Kelleher, 2016). Conversational participants employ such mechanisms to achieve a mutual understanding of the scene. We see this as a promising line of our future work.

References

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

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge.

Stephen Clark. 2015. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics — second edition*, chapter 16, pages 493–522. Wiley – Blackwell.

Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the apprehension of Over, Under, Above and Below. *Journal of Memory and Language*, 44(3):376–398.

Kenny R. Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V. Richards. 2005. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bernhard Nebel, and Thomas Barkowsky, editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, volume 3343 of *Lecture Notes in Computer Science*, pages 98–110. Springer Berlin Heidelberg.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Kees van Deemter. 2016. *Computational models of referring: a study in cognitive science*. The MIT Press, Cambridge, Massachusetts and London, England.

Simon Dobnik and John D. Kelleher. 2016. A model for attention-driven judgements in type theory with records. In Julie Hunter, Mandy Simons, and Matthew Stone, editors, *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue*, volume 20, pages 25–34, New Brunswick, NJ USA, July 16–18.

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, September 4. http://www.dobnik.net/simon/documents/thesis.pdf.

Klaus-Peter Gapp. 1994. A computational model of the basic meanings of graded composite spatial relations in 3d space. In *AGDM*, pages 66–79.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In Hanspeter A. Mallot, editor, *Second International Workshop on Models and Representations in Spatial Cognition (MRSC)*, page 21, Schloss Hohentübingen, Tübingen, Germany, April 6–7. Cognitive Neuroscience, Department of Biology, University of Tübingen.

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

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

Thomas Hörberg. 2008. Influences of form and function on the acceptability of projective prepositions in swedish. *Spatial Cognition & Computation*, 8(3):193–218.

John D. 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 the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 745–752, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simon Kirby, Kenny Smith, and Hannah Cornish. 2008. Language, learning and cultural evolution: How linguistic transmission leads to cumulative adaptation. In Robin Cooper and Ruth Kempson, editors, *Language in Flux: Dialogue Coordination, Language Variation, Change and Evolution*, Communication, Mind and Language. College Publications, London.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*, online:1–35, December 18.

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.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

James Pustejovsky. 1995. *The generative lexicon*. MIT Press, Cambridge, Mass.

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.

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

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1213–1223, Berlin, Germany, August 7–12, 2016. Association for Computational Linguistics.

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

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.