KILLE: a Framework for Situated Agents for Learning Language Through Interaction

Simon Dobnik
CLASP
University of Gothenburg, Sweden
simon.dobnik@gu.se

Erik Wouter de Graaf
University of Gothenburg, Sweden
kille@masx.nl

Abstract

We present KILLE, a framework for situated agents for learning language through interaction with its environment (perception) and with a human tutor (dialogue). We provide proof-of-concept evaluations of the usability of the system in two domains: learning of object categories and learning of spatial relations.

1 Introduction

Situated conversational robots have to be capable of both linguistic interaction with humans and interaction with their environment through perception and action if we want them a part of our daily lives. Humans interact through language very efficiently and naturally and since most of them are not expert programmers, interaction with a robot in a natural language will be preferred. Secondly, by being part of the human environment containing everyday objects such as tables and chairs, robots too have to have knowledge how humans structure and organise their world which is again reflected in human language.

Connecting language with perception and action is commonly known as grounding (Harnad, 1990; Roy, 2002). The main challenge in grounding is that we are connecting two representation systems, (i) a perceptual which is commonly captured in physical sciences as continuous real-valued features and (ii) a symbolic conceptual system that makes up human language. There is no one-to-one correspondence between the two: linguistic descriptions such as "close to the table" and "red" correspond to some function predicting the degree of acceptability over physical or colour space (Logan and Sadler, 1996; Roy, 2002; Roy, 2005; Skočaj et al., 2010; Matuszek et al., 2012; Kennington and Schlangen, 2015; McMahan and Stone, 2015). The relations between concepts are not flat but are made increasingly more abstract, structures are embedded and recursive (Fellbaum, 1998; Tenenbaum et al., 2011) and organised at several representational layers (Kruijff et al., 2007). It follows that several descriptions may be equally applicable for the same situation: the chair can be "close to the table" or "to the left of the table" which means *vagueness* is prevalent in grounding. This however, can be resolved through interaction by adopting appropriate interaction strategies (Kelleher et al., 2005; Skantze et al., 2014; Dobnik et al., 2015).

The meaning of words is not grounded just in perception and action but also grounded in particular linguistic interactions or conversations: participants continuously adapt and agree on the meaning of words as a part of their interaction (Clark, 1996; Fernández et al., 2011). This means that having a static model of grounded language which is learned offline from a corpus with a situated robot is not enough but this must be continuously adapted as the interaction unfolds (Skočaj et al., 2011; Matuszek et al., 2012). The idea of dynamic, continuously updated grounded language models is parallel to dynamic, continuously updated maps of the environment that have been commonly used in mobile robotics for a while (Dissanayake et al., 2001). Static models used in early robotics (Winograd, 1976) were just not able to deal with any changes in its environment and the uncertainty that these bring. We want to take the same view for language which is dynamically adjusted through interaction strategies.

In this paper we describe a framework for situated agents that learn grounded language incrementally and online called KILLE (Kinect Is Learning LanguagE) with the help of a human tutor in the fashion previously described. KILLE is a non-mobile table-top robot connecting Kinect sensors with image processing and classification and a spoken dialogue system. The system learns to recognise objects presented to it by a human tu-

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 162–171,
Gothenburg, Sweden, 23-24 May 2017.

tor from scratch. It can direct learning by asking for more objects of a particular category if it is not able to classify them with sufficient reliability. If more objects of a particular category are available in the scene and the system is able to recognise them, the system queries the user to describe spatial relations between them. Each of these kinds of descriptions focus on different perceptual features and represent two fundamental linguistic semantic categories: entities and entity relations. Overall, KILLE combines both passive and active learning which is incremental at the level of both kinds of linguistic categories.

The contributions of the KILLE framework are two-fold: (i) from the computational perspective it provides a platform for building models of situated language learning and answering questions how to integrate and test existing language technology tools (primarily intended for processing corpora) in an interactive tutoring framework; (ii) it also provides a platform for testing linguistic and psycho-linguistic theories, formalisms and applications on grounding language in interaction (Larsson, 2013; Dobnik et al., 2013) and implementing them computationally. This paper focuses on the construction of the Kille framework and its properties while it also provides a proof-of-concept evaluation of such learning of simple object and spatial relations representations. The paper is organised as follows. In Section 2 we describe the main components of the system. In Sections 3 and 4 we describe the perceptual representations and dialogue strategies that have been implemented so far. Section 5 describes proof-concept learning of objects and Section 5.2 learning of spatial descriptions that demonstrate the usability of the framework. We give conclusions and discussion of future work in Section 6.

2 The KILLE system

The system and the architecture that KILLE is using are similar to two existing systems for incremental interaction (Schlangen and Skantze, 2011) IrisTK (Skantze and Al Moubayed, 2012) and InproTK (Kennington et al., 2014). The difference is that instead of starting from the perspective of incremental language processing and dialogue management we focus on the mapping between language and robot's sensors and actuators and how to learn such mappings through particular dialogue and interactional strategies. Therefore, we opt for a Robot Operating System (ROS) (Quigley et al., 2009) as our middle-ware which provides a common framework for building modules that communicate with each other (send and receive information) and runs on a variety of popular robotic hardware implementations which makes modules portable between them. We work with a very simple robotic hardware: a Microsoft Kinect sensor supported by the *libfreenect* library integrated in ROS.[1] The Kinect provides us with three sensors: an RGB camera of resolution 640x480, a depth sensor, which a structured-light 3D scanner that can perceive in a distance between 70 and 300 cm and gives a 3d representation of object, and a multi-array microphone (not used). The Kinect sensor is attached to a laptop computer running ROS and other software and together they represent our interactive robot. This robot does not have actuators which means it is not mobile and so it cannot turn gaze and focus on objects not in its vision field. Objects have to be brought into its attention field. We use glass pedestals as object support as to avoid the problem of object segmentation (e.g. object and hand) as glass is not detected by the depth sensor. Although simple, the platform satisfies well our requirements for incremental situated learning of perceptual language through dialogue. Through ROS the system can be ported to other, more sophisticated robotic platforms with very little modification.

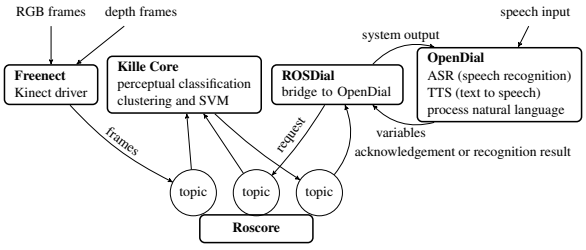

Figure 1: Kille modules

Figure 1 shows the main software modules that make up Kille and how they communicate with each other within the ROS framework. Each component (e.g. Kille, ROSDial, etc.) is a node which may have one or several topics. A node can publish to a topic or subscribe to a topic from another node. *Roscore* is a special node which is responsible for communication. As communication is per-

[1] See http://wiki.ros.org/freenect_stack We use an older version of Kinect hardware 1414 which is better supported by *libfreenect*.

formed over TCP/IP nodes can be distributed over several machines.

For dialogue management we use OpenDial[2] (Lison, 2013) which is a domain independent dialogue manager supporting probabilistic rules. It comes pre-packaged with several other popular NLP tools and interfaces to ASR and TTS systems. User utterances are ran through ASR and POS-tagged with the MaltParser. The output is then processed by a series of dialogue rules which define pre-conditions and post-conditions of their application. Since this is a perceptual dialogue system, dialogue rules involve both linguistic information and information received from the perceptual module of the system (Kille Core), for example the names of the objects detected and the certainty of detection, the spatial relation between them, etc. The dialogue rules can define further dialogue moves or actions for the perceptual system to take. In order to use OpenDial in our ROS configuration we had to build a bridge between the two which we call ROSDial. This sets up a ROS node and an instance of OpenDial. As OpenDial, ROSDial is written in Java. ROSDial translates the messages between OpenDial's information state and a ROS topic. It also ensures that Kille and OpenDial are synchronised. As the interaction is driven by OpenDIAL, sending requests to Kille is straightforward. However, it can also happen that a perceptual event is detected by Kille which the dialogue manager should act upon. ROSDial periodically instantiates a dialogue rule to interpret for any new information that has been pushed to its information state from Kille Core. Finally, Kille Core is written in Python and handles all perceptual representations and learning. The representations of objects and spatial relations can be saved and reloaded between sessions. Kille Core also sends and receives messages both to and from the Kinect library, e.g. scanned perceptual data, and ROSDial, e.g. linguistic data. Both ROSDial and Kille Core are available on Github.[3]

3 Perceptual representations

For visual representations we use OpenCV (Open Source Computer Vision)[4] which is a popular library for computer vision including machine learning applications. It is natively written in C and C++, but has interfaces for other languages including Python (Bradski and Kaehler, 2008). It is also optimised for real-time applications. Through ROS we receive real-time frames from Kinect which include both data from the depth sensor and the visual RGB sensor. The frames are converted to the OpenCV format which is compatible with *NumPy* arrays and which allows for fast computational manipulation.

The visual processing is performed in two steps. In the first step the information from the depth sensor is used to detect the object in focus and remove irrelevant foreground and background in the RGB image. The depth sensor of Kinect cannot detect objects that are closer to it than 70 cm (Figure 2a). We define background as anything that is further away than 100 cm from the Kinect sensor and remove it (Figure 2b). This leaves us 30 centimetre of space that we can present objects in, which turns out to be sufficient and works well for our experiments.

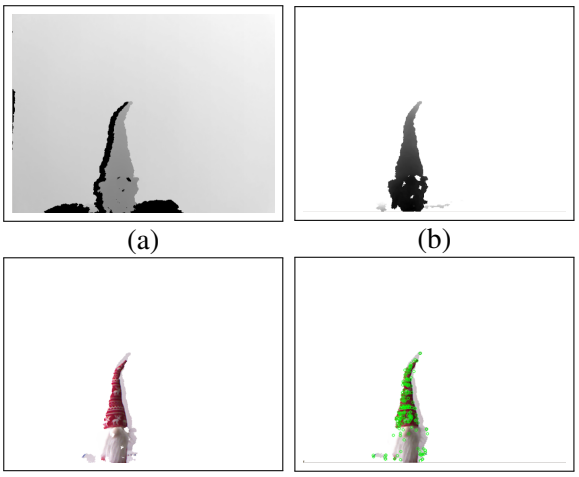

Figure 2: A perception of a plush gnome from the depth sensor (a) including the background, (b) with the background removed, (c) with the RGB image superimposed, and (d) with SIFT features detected in the image. The black border in (a) is a perceptual artefact arising from the interference of sensors.

In the next step, the RGB image is processed and only those pixels are preserved in the image that correspond to the remaining depth points (Figure 2c). These pixels, representing the object in focus, are processed with the SIFT (Scale-Invariant Feature Transform) algorithm (Lowe, 1999) which detects points of interest in a still image (Fig-

[2]https://github.com/plison/opendial
[3]https://github.com/masx/Kille
[4]http://opencv.org

ure 2c). Each point is represented as descriptor, a high dimensional vector which contains information about the visual properties surrounding each key-point, its orientation and location. SIFT is intended to be scale-invariant which means it is robust against changes that are introduced by scaling, can handle rotation and small changes in illumination. Note that objects (including those that are very similar and belong to the same category) will have different numbers of SIFT descriptors detected depending on their visual properties: some objects have a lot of visual details while others are smoother. In our interactive scenario this will also be the case each time when the same object will be re-introduced to the system as the captured frame will be slightly different from the previously captured one because of slight changes in location, lighting and camera noise.

4 Dialogue games

In the dialogue manager we implemented dialogue rules that act upon both linguistic and perceptual knowledge and allow us to have the following conversational games (Kowtko et al., 1992) with users.

1. **Direct instruction: objects**

   User: This is a cup.
   System: [Object is focused on and observed.]
   S: OK, I learned a cup.

2. **Direct instruction: spatial relations**

   U: The book is to the right of the mug.
   S: [Objects are focused on and observed.]
   S: OK. This relation is called "to the right of".
   S: [Objects are focused on and observed.]
   U: Where are the objects?
   S: The book is to the right of the mug.

3. **Describing objects and relations with (un)certainty**

   U: What is this?
   S: [Object is focused on and observed.]
   U: The object is thought to be a book, but it might also be a mug.

4. **Reinforcement from feedback**

   U: What is this?
   S: [Object is focused on and observed.]
   S: This is a cup.
   U: That is correct.
   S: [Observation is added to the models.]

5. **Self initiative: spatial relations**

   S: [Objects are focused on and observed.]
   S: One of the objects is book, the other one is box. Please tell me where things are.
   U: The book is to the left of the box.
   S: OK, this relation is called "to the left of".

6. **Unlearning immediately (e.g. after a mishearing)**

   U: This is a book.
   S: [Object is focused on and observed.]
   S: OK, this is a box.
   U: That is not what I said.
   S: OK, sorry. I will ignore it.

7. **Unlearning later**

   U: Forget a book.
   S: OK. I forgot a book.

8. **Renaming an object**

   U: This is a book.
   S: [Object is focused on and observed.]
   S: OK, this is a box.
   U: I said a book.
   S: OK, sorry. This is a book.

5 Interactive perceptional learning

In the following subsections we present a proof-of concept implementation and evaluation of perceptual learning through interaction which demonstrates the usability of the Kille framework.

5.1 Learning to recognise objects

As described in Section 3 every scan, even of the same scene, gives us a different set of SIFT descriptors which as data-structures are multidimensional vectors. One approach to use SIFT descriptors in classification is to pre-cluster them to bags of visual words (BoVW) (see (Bruni et al., 2014) for discussion) and then apply their occurrence counts as features in a classifier such as Linear Support Vector Machine (SVM). However, in the current implementation we chose a simple lazy-learning method based on SIFT clustering as it better fits with the incremental learning scenario. With the SVM method we would have to store and cluster instances and re-train the model on each instance update, thus doing far more computational work. Since the domain and the number of examples are small, lazy-learning is justified.

The SIFT descriptors of each object instance are stored in a database and then at each classification step the current SIFT descriptions are compared against objects stored in memory and the category of the best matching object is returned. The matching of SIFT descriptors as k-nearest neighbours has been implemented in the FLANN library (Muja and Lowe, 2009). This takes the longest list of descriptors (either from the current instance or

a database instance) and matches each descriptor to k descriptors in the other list (in our case $k = 2$). The matched tuples (some of which will have zero or no similarity) have to be filtered. For this the ratio test of (Lowe, 2004) is used which calculates the Euclidean distance between two descriptors and those pairs that fall below the empirically defined threshold of 0.75 are discarded. Since different object representations contain a different number of SIFT features and there is a bias that representations with a small number of features match representations with a large number of features, we take the harmonic mean of the ratio $\frac{\#Matched}{\#Model}$ and the ratio $\frac{\#Matched}{\#Perceived}$ as the final matching score. In the evaluation 10 consecutive scans are taken and their recognition scores are averaged to a single score. This improves the performance as it makes observations more stable to noise but decreases the processing speed. The name of the item in the database with the closest match is returned. If there are several top-ranking candidates of the same category, their category is returned with their mean recognition score.

The location of the recognised object is estimated by taking the locations of the twenty matched descriptors with the shortest distance.

To evaluate the system's performance in an interactive tutoring scenario we chose typical household objects that could be detected in the perception field of the Kinect sensor and which fall into the following 10 categories: apple, banana, teddy bear, book, cap, car, cup, can of paint, shoe and shoe-box. A human tutor successively reintroduces the same 10 objects to the system in a pre-defined order over four rounds trying to keep the presentation identical as much as possible. In each round all objects are first learned and then queried. To avoid ASR errors both in learning and generation text input is used.

The average recognition scores over four rounds are shown in Table 1. We choose the name of the object with the highest recognition score. The highest values follow all but in one case the diagonal which means that on overall objects are recognised correctly. The only problematic object is the cap which has been consistently confused with a banana. SIFT features do not contain colour information according to which these two categories of objects could be distinguished. There were a few individual confusions which did not affect the overall mean score (not shown in Table 1): the shoe-box was confused with a car in rounds 1 and 2, and with an apple in round 3. The apple was recognised as a banana in round 2. Otherwise, the system performed extremely accurately. The last column C-NI gives *Correct-NextIncorrect* score (a difference between the matching score of the target object with the object of the correct category and the matching score of the target object with the closest matching object not of the correct category) which shows on average how visually distinct the object is. The models of most objects preserve significant distinctiveness over presentations and learning across the 4 rounds. If we rank objects by this score, we get the following ranking (from more distinct to least distinct): book $>$ car $>$ shoe $>$ cup $>$ banana $>$ bear $>$ apple $>$ paint $>$ shoe-box $>$ cap.

In the second experiment we evaluated re-recognition of objects at different degrees of rotation (45°, 90°, 135°, 180°, 225°, 270° and 315° in the clockwise direction) using the final model for objects built after the completion of round 4 in the previous experiment. Already at a rotation of 45° 6 out of 10 objects are mis-recognised. Objects are affected by rotation in different ways since their sides are visually distinct. For example shoe-box and to a large extent apple are correctly classified at most rotational angles. We observe similar classification scores at those angles at which symmetric sides of objects are exposed: 45°:315° and 90°:270°, 135°:225° and 0°:180°.

In the third experiment we tested the model from the end of the round 4 in the first experiment on 3 new objects of each category. The models for experiment one did not extend well to different objects for most categories. Only apples (2 out of 3 new objects) and shoe-boxes (all 3 new objects) were recognised correctly. Shoe-box is also the most common mis-classification which means it is similar to other objects.

Overall, the results and the discussion in this section show that our system is able to learn to recognise objects incrementally through interaction with a human tutor from just a few observations. The testing of our models in new contexts from the context in which they were learned (rotation and classification of different objects of the same category) demonstrate how sensitive our models are to the changes of contexts which are likely to arise in the interactive scenario. Of course, learning observations of objects in these

→	apple	banana	bear	book	cap	car	cup	paint	shoe	shoe-box	C-NI
apple	**.343**	.227	.076	.046	.099	.058	.126	.074	.053	.166	.116
banana	.201	**.357**	.058	.035	.085	.087	.148	.066	.046	.124	.155
bear	.080	.121	**.260**	.074	.089	.091	.120	.099	.074	.136	.123
book	.142	.233	.074	**.496**	.114	.197	.246	.130	.085	.220	.250
cap	.122	**.208**	.076	.049	.146	.096	.103	.083	.061	.114	-.062
car	.104	.183	.053	.067	.077	**.414**	.119	.076	.069	.149	.231
cup	.099	.145	.063	.066	.091	.052	**.330**	.094	.054	.120	.185
paint	.119	.140	.075	.076	.083	.147	.121	**.221**	.062	.111	.075
shoe	.078	.123	.070	.056	.079	.116	.124	.076	**.319**	.103	.196
shoe-box	.190	.332	.099	.188	.145	.305	.313	.166	.111	**.376**	.044

Table 1: Average recognition scores over four rounds. The object tested are represented in rows. Columns indicate the categories that they were recognised as.

contexts specifically would increase the success of object recognition. Thus, the experiments also point to the complexity of the object recognition.

5.2 Learning to recognise spatial relations

Once the system learns to detect several objects in the scene it starts querying the user to describe spatial relations between them. The semantics of spatial relations requires at least three components of meaning: (i) knowledge about the geometrical arrangement of objects in the scene; (ii) world knowledge about the objects involved in particular how they interact with each other or what is our take on the scene; (iii) dialogue interaction between conversational partners, for example in coordinating and negotiation perspective or the origin of the frame of reference (FoR) used. We hope that Kille will provide us a platform for modelling and testing the interaction between all three components, some of which, for example (ii), are learned from a large corpus off-line. Here we mainly focus on interactive learning of the geometric component (i).

First, the system must recognise the target and the landmark objects ("the gnome/TARGET is to the left of the book/LANDMARK") both in the linguistic string and the perceptual scene. Twenty highest ranking SIFT features are taken for each object and their x (width), y (height) and z (depth) coordinates are averaged, thus giving us the centroid of the 20 most salient features of an object.[5] We chose the number 20 based on practical experience. Higher numbers of features are more demanding for processing in real time. The origin of the coordinate frame for spatial templates must be at the centre of the landmark object which means that the coordinates of the target must be expressed as relative to the landmark's location. A further transformation of the coordinate frame could be made depending on the orientation of the viewpoint that sets the perspective. However, in our scenario the geometric coordinate frame was always relative to the orientation of Kille. Of course, in conversations with Kille humans could describe locations from a different perspective which means that this can lead (in an absence of a model of FoR assignment) to a more complicated/noisy and ambiguous model of geometric spatial template learned. For example the same region could be described as "to the left of" and "to the right of". The relativised location of the target to the landmark are fed to a Linear Support Vector Classifier (SVC) with descriptions as target classes.

A human tutor taught the system by presenting it the target object (a book) at 16 different locations in relation to the landmark (the car) as shown in Figure 3. The locations were arranged so that there were 8 locations separated at 45° at two different distances around and from the landmark. These objects were chosen because they have achieved a good recognition accuracy in the previous experiments. The book was shown to the system three times per location in a randomised order which gave us 48 presentations. The target was moved after each presentation. This ensured that there was no semantic influence on descriptions between the presentations.[6] The spatial descriptions that the human instructor used were *to the left of*, *to the right of*, *in front of*, *behind of*, *near* and *close to* (6). The first 4 descriptions are *projective descriptions* and require grounding of FoR, while the last two are *topological descrip-*

[5] This is a simplification as object shape is only partially expressed in the y variable. This way we distinguish between tall and short objects. Note also that the variables x and z describe object location while y describes object property.

[6] In a different evaluation setting we might want to explore such bias.

tions and do not require grounding of the FoR, only distance. The relativised spatial coordinates implicitly encode both of these features.

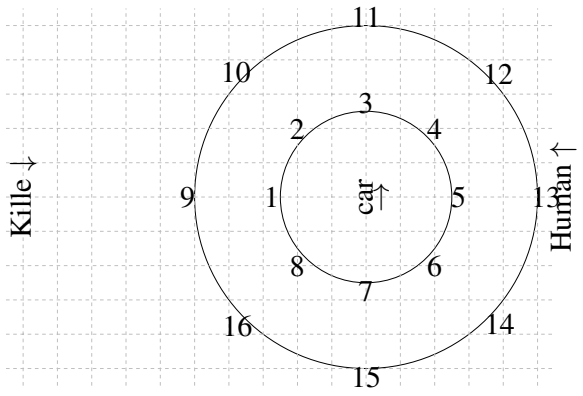

Figure 3: The locations of the target ("book"), the landmark ("car") and the conversational partners (Kille and the human). 1 and 9 have been slightly relaxed, as Kille would not be able to detect the car behind the book otherwise.

Note that spatial descriptions are not mutually exclusive: location 4 in Figure 3 could be described as "near", "close", "to the right of" and "in front of" (taking the human FoR) and "near", "close", "to the left of" and "behind" (taking Kille's FoR) which makes learning a difficult task. In the evaluation we are interested if the system would agree strictly with human observers on the most relevant description for that context, if this is not the case, would the system generate an alternative acceptable description. The agreement between annotators is highly informative as it tells us about the difficulty of the task.

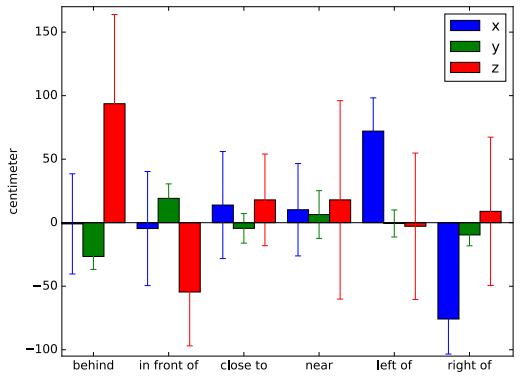

Figure 4: Average x, y and z values for spatial relations.

Figure 4 shows the average values of the x (width), y (height) and z (depth) features with the origin on the landmark object for the instances in our learning dataset. There is quite a clear opposition between "behind" and "in front of" as well as "to the left of" and "to the right of" which means that the instructor was consistent in the usage of the FoR. What is interesting is that different FoR is used for the front-and-back and the lateral dimensions. "behind" and "in front of" are grounded in an FoR relative to Kille (away from and towards Kille respectively) while "left" and "right" are grounded in an FoR relative to the conversational partner (e.g. "to the left of" corresponds to positive x values. Effectively, a split FoR is used. In this scenario, there is no effect of conversational roles on the assignment of FoR (information giver and information receiver) which has been reported by (Schober, 1995). The reason why the target is appears to be lower when it is described to be "behind" the landmark is due to the positioning of the sensor higher than the landmark at an angle looking down. This means that objects further away appear shorter. Finally both "close" and "near" show short distances from the landmark in each dimension as expected. However, there is, at least in this model, no clear difference between these two descriptions.

The performance of the system was independently evaluated by two human conversational partners, one of whom was also the tutor during the learning phase. As during learning, the target object was placed in one of the 16 locations and each location was used twice, which gave each human to evaluate a total of 32 situations which were presented in a random order. After each object placement, the evaluators first independently wrote down the description they would use to describe the scene. Then the system would be queried to describe the location of the target. The system's response was recorded and also whether the evaluators agreed with the generated description or not.

As mentioned earlier, several spatial descriptions may apply to the same location of the target and the landmark. The observed strict agreement between the evaluators independently choosing a description is 0.5313 (they independently choose the same description in just over 1/2 of cases). However, when we correct this value by agreement by chance in the form of the Kappa coefficient (κ), the estimated strict agreement between

the evaluators is $\kappa = 0.4313$. Choosing a spatial description is thus quite a subjective task.

Match	**Evaluator 1**		**Evaluator 2**		**Evaluator 1 + 2**	
Independent	8	0.25	7	0.2188	15	0.2344
Secondary	11	0.3438	13	0.4063	24	0.375
Indep. + Second.	19	0.5938	20	0.6251	39	0.6094
Incorrect	13	0.4063	12	0.375	25	0.3906
Total	32	1	32	1	64	1

Table 2: Observed agreement between the evaluators and the system

The observed agreement between the evaluators and the system is shown in Table 2. The evaluators and the system independently chose the same description in 23.44% of cases which is a decrease from 53.13% where only evaluators are compared with each other. However, even if the description was not the one that evaluators chose in this situation, the evaluators thought that the generated description was nonetheless a good description in further 37.5% of situations. Overall, evaluators were satisfied with 60.94% of generations, while 39.06% were considered incorrect. Given the difficulty of the task and that the system on average had a chance to learn each description only from 8 trials (48/6 = 8), thus not observing each description at all possible 16 locations, the results are encouraging and are close to what has been reported for a similar task in the literature (Dobnik, 2009).

	behind	front	left	right	close	near	Total
behind	**4**	2	1	0	0	2	9
front	0	**5**	3	3	6	0	17
left	0	6	**1**	0	0	0	7
right	4	1	3	**3**	0	1	12
close	1	9	1	0	**1**	2	14
near	1	1	1	0	1	**1**	5
Total	10	24	10	6	8	6	64

Table 3: Agreement between two human evaluators (rows) and the system (columns)

Table 3 shows a confusion matrix for all 64 trials. The Kappa coefficient, thus the strict observed agreement of 0.2344 (Table 2) discounted by the agreement by chance is $\kappa = 0.0537$. If we examine Table 3 we can see, as also shown in Table 2, that non-agreements involve those descriptions that are appropriate alternatives. For example, we expect topological descriptions (e.g. "close") to partially overlap with projective descriptions (e.g. "front") or with other projective descriptions (e.g. "left" and "front"). The data also shows that humans and evaluators have a slightly different preference for assigning descriptions. Humans assign descriptions with the following likelihoods (from the highest to the lowest): "front" (0.2656) > "close" (0.2188) > "right" (0.1875) > "behind" (0.1406) > "left" (0.1094) > "near" (0.0781) while the system has the following preference "front" (0.375) > "behind"/"left" (0.1563) > "close" (0.125) and > "right"/"near" (0.0938). The most important differences are thus in the usage of "close"/"right" *vs* "behind"/"left" and demonstrate the subjective nature of the task and possibly a usage of different FoRs.

6 Conclusion and future work

We presented Kille, a framework for situated agents for learning language through interaction. This is based on a Robotic Operating System (ROS) which simplifies the development of new applications and their communication, as well as allowing the system to be ported to a variety of more sophisticated popular robotic platforms. We focus on the linguistic interactional aspects with a situated agent in the context of learning through instruction and therefore our three main modules are the robotic perceptual system provided by the Kinect sensor, a dialogue system and a module for classification of grounded lexical meanings. We demonstrate and evaluate the usability of the system on two proof-of-concept applications: learning of object names and learning of spatial relations.

As stated in the introduction we hope that the framework will allow us to explore computational and linguistic questions related to situated learning. In particular, we are interested in how (i) different machine learning methods can be used with an interactive tutoring scenario including the application of image convolutions from deep learning to replace SIFT features. (ii) Integration of classifiers learned offline from a large corpus with the interactive learning and classification is also an open question. On the language side we are interested in (iii) what kind of interaction strategies or dialogue games are relevant in this scenario, (iv) how can these games be implemented in a situated dialogue system in terms of dialogue moves operating on linguistic and perceptual representations and linking to machine learning or classification, (v) and how effective individual dialogue games are in respect to the rate of learning.

References

Gary Bradski and Adrian Kaehler. 2008. *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.".

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49(1-47).

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

M. W. M. G Dissanayake, P. M. Newman, H. F. Durrant-Whyte, S. Clark, and M. Csorba. 2001. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotic and Automation*, 17(3):229–241.

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2013. Modelling language, action, and perception in Type Theory with Records. In Denys Duchier and Yannick Parmentier, editors, *Constraint Solving and Language Processing: 7th International Workshop, CSLP 2012, Orléans, France, September 13–14, 2012, Revised Selected Papers*, volume 8114 of *Lecture Notes in Computer Science*, pages 70–91. Springer Berlin Heidelberg.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In Christine Howes and Staffan Larsson, editors, *Proceedings of goDIAL - Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden, 24–26th August.

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.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, Mass.

Raquel Fernández, Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and David Schlangen. 2011. Reciprocal learning via dialogue interaction: Challenges and prospects. In *Proceedings of the IJCAI 2011 Workshop on Agents Learning Interactively from Human Teachers (ALIHT)*, Barcelona, Catalonia, Spain.

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

J.D. Kelleher, F. Costello, and J. van Genabith. 2005. Dynamically structuring updating and interrelating representations of visual and linguistic discourse. *Artificial Intelligence*, 167:62–102.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China, July. Association for Computational Linguistics.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2014. Inprotks: A toolkit for incremental situated processing. *Proceedings of SIGdial 2014: Short Papers*.

Jacqueline C Kowtko, Stephen D Isard, and Gwyneth M Doherty. 1992. Conversational games within dialogue. HCRC research paper RP-31, University of Edinburgh.

Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2007. Situated dialogue and spatial organization: what, where... and why? *International Journal of Advanced Robotic Systems*, 4(1):125–138. Special issue on human and robot interactive communication.

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

Pierre Lison. 2013. *Structured Probabilistic Modelling for Dialogue Management*. Ph.D. thesis, Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Oslo, 30th October.

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.

David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. IEEE.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, Edinburgh, Scotland, June 27th - July 3rd.

Brian McMahan and Matthew Stone. 2015. A bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Marius Muja and David G Lowe. 2009. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331–340):2.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5.

Deb K. Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer speech and language*, 16(3):353–385.

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

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue and discourse*, 2(1):83–111.

Michael F. Schober. 1995. Speakers, addressees, and frames of reference: Whose effort is minimized in conversations about locations? *Discourse Processes*, 20(2):219–247.

Gabriel Skantze and Samer Al Moubayed. 2012. IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 69–76. ACM.

Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. 2014. Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Communication*, 65:50–66.

Danijel Skočaj, Miroslav Janiček, Matej Kristan, Geert-Jan M. Kruijff, Aleš Leonardis, Pierre Lison, Alen Vrečko, and Michael Zillich. 2010. A basic cognitive system for interactive continuous learning of visual concepts. In *ICRA 2010 workshop ICAIR - Interactive Communication for Autonomous Intelligent Robots*, pages 30–36, Anchorage, AK, USA.

Danijel Skočaj, Matej Kristan, Alen Vrečko, Marko Mahnič, Miroslav Janíček, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich, and Kai Zhou. 2011. A system for interactive learning in dialogue with a tutor. In *IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2011*, San Francisco, CA, USA, 25-30 September.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

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