

# Interactively learning visually grounded word meanings from a human tutor

Yanchao Yu

Interaction Lab  
Heriot-Watt University  
y.yu@hw.ac.uk

Arash Eshghi

Interaction Lab  
Heriot-Watt University  
a.eshghi@hw.ac.uk

Oliver Lemon

Interaction Lab  
Heriot-Watt University  
o.lemon@hw.ac.uk

## Abstract

We present a multi-modal dialogue system for interactive learning of perceptually grounded word meanings from a human tutor. The system integrates an incremental, semantic parsing/generation framework - Dynamic Syntax and Type Theory with Records (DS-TTR) - with a set of visual classifiers that are learned throughout the interaction and which ground the meaning representations that it produces. We use this system in interaction with a simulated human tutor to study the effect of different dialogue policies and capabilities on accuracy of learned meanings, learning rates, and efforts/costs to the tutor. We show that the overall performance of the learning agent is affected by (1) who takes initiative in the dialogues; (2) the ability to express/use their confidence level about visual attributes; and (3) the ability to process elliptical as well as incrementally constructed dialogue turns.

## 1 Introduction

Identifying, classifying, and talking about objects or events in the surrounding environment are key capabilities for intelligent, goal-driven systems that interact with other agents and the external world (e.g. robots, smart spaces, and other automated systems). To this end, there has recently been a surge of interest and significant progress made on a variety of related tasks, including generation of Natural Language (NL) descriptions of images, or identifying images based on NL descriptions e.g. (Bruni et al., 2014; Socher et al., 2014). Another strand of work has focused on learning to generate object descriptions and object classification based on low level concepts/features (such as colour, shape and material), enabling systems to identify and describe novel, unseen images

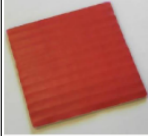

Dialogue Example		Final Semantics in TTR
T: what is this? S: a red circle? T: no, a red square. S: oh, okay.		$X_{=o1} : e$ p2 : red(X) p3 : square(X)
T: what can you see? S: something orange. T: what shape is it S: a square. T: no, it's a circle. S: uhu		$X_{1=o2} : e$ S=s : per p : circle(X1) p1 : orange(X1) p2 : see(S,X1)

Figure 1: Example dialogues

(Farhadi et al., 2009; Silberer and Lapata, 2014; Sun et al., 2013).

Our goal is to build *interactive* systems that can learn grounded word meanings relating to their perceptions of real-world objects – this is different from previous work such as e.g. (Roy, 2002), that learn groundings from descriptions without any interaction, and more recent work using Deep Learning methods (e.g. (Socher et al., 2014)).

Most of these systems using machine learning rely on training data of high quantity with no possibility of online error correction. Furthermore, they are unsuitable for robots and multimodal systems that need to continuously, and incrementally learn from the environment, and may encounter objects they haven't seen in training data. These limitations should be alleviated if systems can learn concepts as and when needed, from situated dialogue with humans. Interaction with human tutors also enables systems to take initiative and seek information they need by e.g. asking questions with the highest information gain (see e.g. (Skocaj et al., 2011), and Fig. 1). For example, a robot could ask questions (Cakmak and Thomaz, 2012) to learn the colour of a “square” or to request to be presented with more “red” things to improve its performance on the concept (see e.g. Fig. 1). Furthermore, such systems could allow for meaning negotiation in the form of clarifica-

tion interactions with the tutor.

This setting means that the system must be *trainable from little data, compositional, adaptive, and able to handle natural human dialogue with all its glorious context-sensitivity and messiness* – for instance so that it can learn visual concepts suitable for specific tasks/domains, or even those specific to a particular user. Interactive systems that learn continuously, and over the long run from humans need to do so *incrementally, quickly, and with minimal effort/cost to human tutors*.

In this paper, we use an implemented dialogue system (see Yu et al. (2016b) and architecture in figure 2) that integrates an incremental, semantic grammar framework, especially suited to dialogue processing – Dynamic Syntax and Type Theory with Records (DS-TTR<sup>1</sup> (Kempson et al., 2001; Eshghi et al., 2012)) with visual classifiers which are learned during the interaction, and which provide perceptual grounding for the basic semantic atoms in the semantic representations (Record Types in TTR) produced by the parser (see Fig. 1).

We use this system in interaction with a simulated human tutor, to test hypotheses about how the accuracy of learned meanings, learning rates, and the overall cost/effort for the human tutor are affected by different dialogue policies and capabilities; specifically: (1) who takes initiative in the dialogues; (2) the agent’s ability to utilise their level of uncertainty about an object’s attributes; and (3) their ability to process elliptical as well as incrementally constructed dialogue turns. The results show that differences along these dimensions have significant impact both on the accuracy of the learned, grounded word meanings, and the processing effort required by the tutors.

## 2 Related work

Please see (Yu et al., 2016b) for a full discussion of related work. Most similar to our work is probably that of Kennington & Schlangen (2015) who learn a mapping between individual words - rather than logical atoms - and low-level visual features (e.g. colour-values) directly. The system is compositional, yet does not use a grammar (the compositions are defined by hand). Further, the groundings are learned from pairings of object references in NL and images rather than from dialogue.

What sets our approach apart from others is: a) that we use a domain-general, incremental se-

mantic grammar with principled mechanisms for parsing and generation; b) Given DS model of dialogue (Eshghi et al., 2015), representations are constructed jointly and interactively by the tutor and system over the course of several turns (see Fig. 1); c) perception and NL-semantics are modelled in a single logical formalism (TTR); d) we effectively induce an ontology of atomic types in TTR, which can be combined in arbitrarily complex ways for generation of complex descriptions of arbitrarily complex visual scenes (see e.g. (Dobnik et al., 2012) and compare this with (Kennington and Schlangen, 2015), who do not use a grammar and therefore do not have logical structure over grounded meanings).

## 3 Experimental Setup

Our goal in this paper is an experimental study of the effect of different dialogue policies and capabilities on the overall performance of the learning agent, which, as we describe below is a measure that combines accuracy of learned meanings with the cost of tutoring over time.

**Design.** We use the dialogue system outlined above to carry out our main experiment with a  $2 \times 2 \times 2$  factorial design, i.e. with three factors each with two levels. Together, these factors determine the learner’s dialogue behaviour: (1) **Initiative (Learner/Tutor)**: determines who takes initiative in the dialogues. When the tutor takes initiative, s/he is the one that drives the conversation forward, by asking questions to the learner (e.g. “What colour is this?” or “So this is a ...” ) or making a statement about the attributes of the object. On the other hand, when the learner has initiative, it makes statements, asks questions, initiates topics etc. (2) **Uncertainty (+UC/-UC)**: determines whether the learner takes into account, in its dialogue behaviour, its own subjective confidence about the attributes of the presented object. The confidence is the probability assigned by any of its attribute classifiers of the object being a positive instance of an attribute (e.g. ‘red’) - see below for how a confidence threshold is used here. In +UC, the agent will not ask a question if it is confident about the answer, and it will hedge the answer to a tutor question if it is not confident, e.g. “T: What is this? L: errm, maybe a square?”. In -UC, the agent always takes itself to know the attributes of the given object (as given by its currently trained

<sup>1</sup>Download from <http://dylan.sourceforge.net>

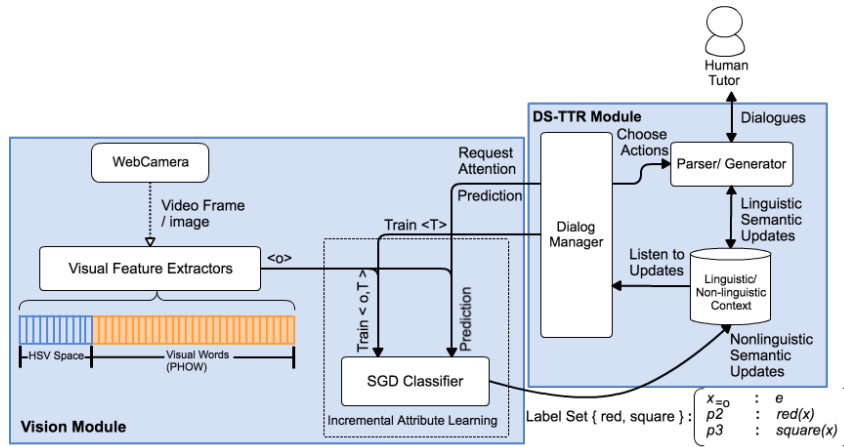


Figure 2: Architecture of the teachable system, see (Yu et al., 2016b)

<b>T+UC+CD</b>	<b>T-UC-CD</b>	<b>T+UC-CD</b>	<b>L+UC+CD</b>
T: This is a ...	T: What (shape) is this?	T: What is this?	L: What colour is this?
L: Errm, a square?	L: This is a circle.	L: Sorry, I don't know.	T: Red.
T: Yes.	T: Yes. What colour is it?	T: Okay, it is a square.	L: Okay.
L: What colour is it?	L: It is red.	L: Okay, I see.	L: Is this a square?
T: Red.	T: No, it's purple.	T: What colour is it?	T: No, a circle.
T: No, it's green.	L: Okay, I see.	L: Is it blue?	L: Okay.
L: Okay, thanks.		T: Yes.	

Figure 3: Example dialogues in different conditions

classifiers), and behaves according to that assumption. (3) **Context-Dependency (+CD/-CD)**: determines whether the learner can process (produce/parse) context-dependent expressions such as short answers and incrementally constructed turns, e.g. “T: What is this? L: a square”, or “T: So this one is ...? L: red/a circle”. This setting can be turned off/on in the DS-TTR dialogue model.

**Tutor Simulation and Policy:** To run our experiment on a large-scale, we have hand-crafted an *Interactive Tutoring Simulator*, which simulates the behaviour of a human tutor. The tutor policy is kept constant across all conditions. Its policy is that of an always *truthful*, *helpful* and *omniscient* one: it (1) has complete access to the labels of each object; and (2) always acts as the context of the dialogue dictates: answers any question asked, confirms or rejects when the learner describes an object; and (3) always corrects the learner when it describes an object erroneously.

**Confidence Threshold:** To determine when and how the agent properly copes with its attribute-based predictions, we use confidence-score thresholds. It consists of two values, a base threshold (e.g. 0.5) and a positive threshold (e.g. 0.9).

If the confidences of all classifiers are under the base threshold (i.e. the learner has no attribute la-

bel that it is confident about), the agent will ask for information directly from the tutor via questions (e.g. “L: what is this?”).

On the other hand, if one or more classifiers score above the base threshold, then the positive threshold is used to judge to what extent the agent trusts its prediction or not. If the confidence score of a classifier is between the positive and base thresholds, the learner is not very confident about its knowledge, and will check with the tutor, e.g. “L: is this red?”. However, if the confidence score of a classifier is above the positive threshold, the learner is confident enough in its knowledge not to bother verifying it with the tutor. This will lead to less effort needed from the tutor as the learner becomes more confident about its knowledge.

However, since a learner with high confidence will not ask for assistance from the tutor, a low positive threshold may reduce the chances that allow the tutor to correct the learner’s mistakes. Hence, we set up an auxiliary experiment, in which we kept all other conditions constant (i.e. assume that the learner has initiative (**L**) and always considers the prediction confidence(**+U**)), but only varied the threshold values. This additional experiment determined a 0.5 base threshold and a 0.9 positive threshold as the most appropriate values for an interactive learning process - i.e.

Table 1: Recognition Score Table

	Yes	LowYes	LowNo	No
Yes	1	0.5	-0.5	-1
No	-1	-0.5	0.5	1

this preserved good classifier accuracy while not requiring much effort from the tutor.

**Recognition score:** We follow metrics proposed by Skocaj et al. (2009). ‘Recognition score’ measures the overall accuracy of the learned word meanings / classifiers, which “rewards successful classifications (i.e. true positives and true negatives) and penalizes incorrect predictions (i.e. false positives and false negatives)” (Skočaj et al., 2009). As the proposed system considers both correctness of predicted labels and prediction confidence on learning tasks, the measure will also take the true labels with lower confidence into account, as shown in Table 1; “LowYes” means that the system made positive predictions but with lower confidence. In this case, the system can generate a polar question to request tutor feedback. “LowNo” is similar to “LowYes”, but for negative predictions.

**Cost:** This measure reflects the effort needed by a human tutor in interacting with the system. Skocaj et. al. (2009) point out that a teachable system should learn as autonomously as possible, rather than involving the human tutor too frequently. There are several possible costs that the tutor might incur, see Table 2:  $C_{inf}$  refers to the cost of the tutor providing information on a single attribute (e.g. “this is red” or “this is a square”);  $C_{ack}$  is the cost for a simple confirmation (like “yes”, “right”) or rejection (such as “no”);  $C_{crt}$  is the cost of correction for a single concept (e.g. “no, it is blue”). We associate a higher cost with correction of statements than that of polar questions. This is to penalise the learning agent when it confidently makes a false statement – thereby incorporating an aspect of trust in the metric (humans will not trust systems which confidently make false statements). And finally, parsing ( $C_{parse}$ ) as well as production ( $C_{production}$ ) costs for tutor are taken into account: each single word costs 0.5 when parsed by the tutor, and 1 if generated (production costs twice as much as parsing).

**Performance Score:** As mentioned above, an efficient learner dialogue policy should consider both classification accuracy (Recognition score)

Table 2: Tutoring Cost Table

$C_{inf}$	$C_{ack}$	$C_{crt}$	$C_{parsing}$	$C_{production}$
1	0.25	1	0.5	1

and tutor effort (Cost). We thus define an integrated measure – the *Overall Performance Ratio* ( $R_{perf}$ ) – that we use to compare the learner’s overall performance across the different conditions:

$$R_{perf} = \frac{\Delta S_{recog}}{C_{tutor}}$$

i.e. the increase in Recognition Score ( $S_{recog}$ ) per unit of the cost, or equivalently the gradient of the curve in Fig. 4c. We seek dialogue strategies that maximise this.

### 3.1 Evaluation and Cross-validation

We performed a 20-fold cross validation with 500 images for training and 100 for testing (see (Yu et al., 2016b) for details of the dataset). For each training instance, the learning system interacts (only through dialogue) with the simulated tutor. Each interaction episode ends either when both the shape and the colour of the object are agreed upon, or when the learner requests to be presented with the next image. We define a learning step as comprised of 10 such episodes. At the end of each learning step, the system is tested using the test set. The values used for the Tutoring Cost and the Recognition Score at each learning step correspond to averages across the 20 folds.

## 4 Results

Fig. 3 shows example interactions between the learner and the tutor in some of the experimental conditions. Note how the system is able to deal with (parse and generate) utterance continuations as in  $T+UC+CD$ , short answers as in  $L+UC+CD$ , and polar answers as in  $T + UC + CD$ .

Fig. 4 plots Recognition Score against Tutoring Cost directly. Note that it is expected that the curves should not terminate in the same place on the x-axis since the different conditions incur different total costs for the tutor. The gradient of this curve corresponds to *increase in Recognition Score per unit of the Tutoring Cost*. It is the gradient of the line drawn from the beginning to the end of each curve ( $\tan(\beta)$  on Fig. 4) that constitutes our main evaluation measure of the system’s overall performance in each condition, and it is this measure for which we report statistical significance re-

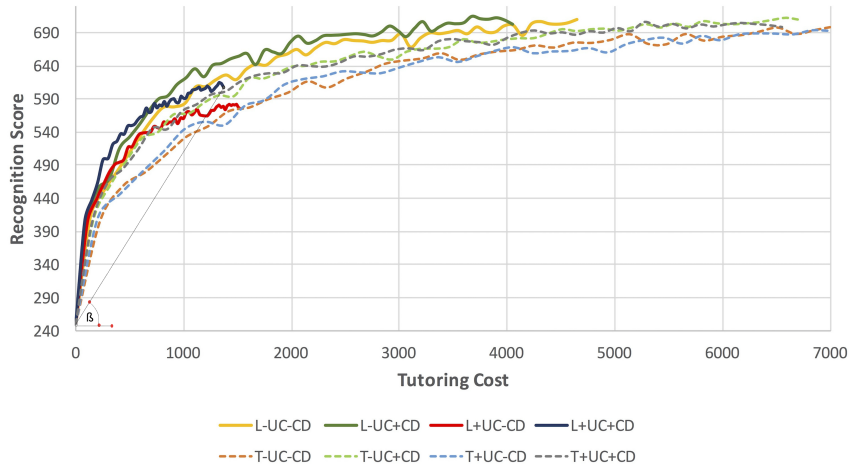


Figure 4: Evolution of Overall Learning Performance

sults: a between-subjects ANOVA shows significant main effects of Initiative ( $p < 0.01$ ;  $F = 469.2$ ), Uncertainty ( $p < 0.01$ ;  $F = 179.8$ ) and Context-Dependency ( $p < 0.01$ ;  $F = 20.12$ ) on the system’s overall performance. There is also a significant Initiative $\times$ Uncertainty interaction ( $p < 0.01$ ;  $F = 181.72$ ).

## 5 Discussion

The cumulative cost for the tutor progresses more slowly when the learner has initiative (L) and takes its confidence into account in its behaviour (+UC). This is so because *a form of active learning* is taking place here: the learner only asks a question about attribute if it isn’t confident enough already about that attribute. As the agent is exposed to more training instances its subjective confidence about its own predictions increases over time, and thus there is progressively less need for tutoring. On the other hand, the Recognition Score increases more slowly too in the L+UC conditions. This is because the agent’s confidence score in the beginning is unreliable as it has only seen a few training instances: in many cases it doesn’t have any interaction with the tutor and so there are informative examples that it doesn’t get exposed to.

However, comparing the gradients of the two curves on Fig. 4 shows that the above trade-off between Recognition Score and Cost is in fact a good one: the overall performance of the agent is significantly better in the L+UC conditions (recall the Initiative  $\times$  Uncertainty interaction). The significant main effect of Context-Dependency on overall performance is explained by the fact that

in +CD conditions, the agent can process context-dependent and incrementally constructed turns, leading to less repetition, shorter dialogues, and so better overall performance.

## 6 Conclusion and Future work

We have presented a multi-modal dialogue system that learns grounded word meanings from a human tutor, incrementally, over time. The system integrates a semantic grammar for dialogue (DS), a logical theory of types (TTR), with a set of visual classifiers in which the TTR semantic representations are grounded. We used this implemented system to study the effect of different dialogue policies and capabilities on the overall performance of a learning agent - a combined measure of accuracy and cost. The results show that in order to maximise its performance, the agent needs to take initiative in the dialogues, take into account its confidence about its predictions, and be able to process natural, human-like dialogue. Ongoing work uses Reinforcement Learning to acquire adaptive dialogue policies that optimise such an agent’s performance (Yu et al., 2016a).

## Acknowledgements

This research is supported by the EPSRC, under grant number EP/M01553X/1 (BABBLE project), and by the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 688147 (MuMMER<sup>2</sup>).

<sup>2</sup><http://mummer-project.eu/>

## References

- Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res. (JAIR)*, 49(1–47).
- Maya Cakmak and Andrea Thomaz. 2012. Designing robot learners that ask good questions. In *Proc. HRI*.
- Simon Dobnik, Robin Cooper, and Staffan Larsson. 2012. Modelling language, action, and perception in type theory with records. In *Proceedings of the 7th International Workshop on Constraint Solving and Language Processing (CSLP’12)*, pages 51–63.
- Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. 2012. Conversational interactions: Capturing dialogue dynamics. In S. Larsson and L. Borin, editors, *From Quantification to Conversation: Festschrift for Robin Cooper on the occasion of his 65th birthday*, volume 19 of *Tributes*, pages 325–349. College Publications, London.
- A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, London, UK. Association for Computational Linguistics.
- Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax: The Flow of Language Understanding*. Blackwell.
- Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the Conference for the Association for Computational Linguistics (ACL-IJCNLP)*. Association for Computational Linguistics.
- Deb Roy. 2002. A trainable visually-grounded spoken language generation system. In *Proceedings of the International Conference of Spoken Language Processing*.
- Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 721–732, Baltimore, Maryland, June. Association for Computational Linguistics.
- Danijel Skočaj, Matej Kristan, Alen Vrečko, Marko Mahnič, Miroslav Jančiček, 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 *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2011, San Francisco, CA, USA, September 25-30, 2011*, pages 3387–3394.
- Danijel Skočaj, Matej Kristan, and Aleš Leonardis. 2009. Formalization of different learning strategies in a continuous learning framework. In *Proceedings of the Ninth International Conference on Epigenetic Robotics; Modeling Cognitive Development in Robotic Systems*, pages 153–160. Lund University Cognitive Studies.
- Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.
- Yuyin Sun, Liefeng Bo, and Dieter Fox. 2013. Attribute based object identification. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 2096–2103. IEEE.
- Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016a. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *(under review)*.
- Yanchao Yu, Oliver Lemon, and Arash Eshghi. 2016b. Comparing dialogue strategies for learning grounded language from human tutor. In *Proc. SEMDIAL 2016*.