The BURCHAK corpus: a Challenge Data Set for Interactive Learning of Visually Grounded Word Meanings

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

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

Gregory Mills
University of Groningen
g.j.mills@rug.nl

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

Abstract

We motivate and describe a new freely available human-human dialogue data set for interactive learning of visually grounded word meanings through ostensive definition by a tutor to a learner. The data has been collected using a novel, character-by-character variant of *the DiET chat tool* (Healey et al., 2003; Mills and Healey, submitted) with a novel task, where a Learner needs to learn invented visual attribute words (such as "burchak" for square) from a tutor. As such, the text-based interactions closely resemble face-to-face conversation and thus contain many of the linguistic phenomena encountered in natural, spontaneous dialogue. These include self- and other-correction, mid-sentence continuations, interruptions, overlaps, fillers, and hedges. We also present a generic n-gram framework for building user (i.e. tutor) simulations from this type of incremental data, which is freely available to researchers. We show that the simulations produce outputs that are similar to the original data (e.g. 78% turn match similarity). Finally, we train and evaluate a Reinforcement Learning dialogue control agent for learning visually grounded word meanings, trained from the BURCHAK corpus. The learned policy shows comparable performance to a rule-based system built previously.

1 Introduction

Identifying, classifying, and talking about objects and events in the surrounding environment are key capabilities for intelligent, goal-driven systems that interact with other humans and the exter-

T(utor): it is a ... [[sako]] burchak.
L(earner):   [[suzuli?]]
T: no, it's sako
L: okay, i see.

(a) Dialogue Example from the corpus
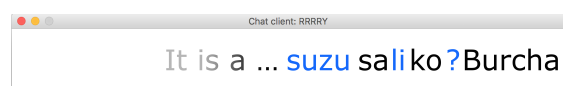
(b) The Chat Tool Window during dialogue in (a) above

Figure 1: Example of turn overlap + subsequent correction in the BURCHAK corpus ('sako' is the invented word for red, 'suzuli' for green and 'burchak' for square)

nal 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 (Bruni et al., 2014; Socher et al., 2014; Farhadi et al., 2009; Silberer and Lapata, 2014; Sun et al., 2013). Another strand of work has focused on incremental reference resolution in a model where word meaning is modeled as classifiers (the so-called Words-As-Classifiers model (Kennington and Schlangen, 2015)).

However, none of this prior work focuses on how concepts/word meanings are *learned and adapted in interactive dialogue* with a human, the most common setting in which robots, home automation devices, smart spaces etc. operate, and, indeed the richest resource that such devices could exploit for adaptation over time to the idiosyncrasies of the language used by their users.

Though recent prior work has focused on the problem of learning visual groundings in interaction with a tutor (see e.g. (Yu et al., 2016b; Yu et

al., 2016a)), it has made use of hand-constructed, synthetic dialogue examples that thus lack in variation, and many of the characteristic, but consequential phenomena observed in naturalistic dialogue (see below). Indeed, to our knowledge, there is no existing data set of real human-human dialogues in this domain, suitable for training multimodal conversational agents that perform the task of *actively learning visual concepts* from a human partner in *natural, spontaneous* dialogue.

(a) Multiple Dialogue Actions in one turn
L: so this shape is wakaki?
T: yes, well done. let's move to the color. So what color is this?

(b) Self-Correction
L: what is this object?
T: this is a sako ... no no ... a suzuli burchak.

(c) Overlapping
T: this color [[is]] ... [[sa]]ko.
L: [[su]]zul[[i?]]
T: no, it's sako.
L: okay.

(d) Continuation
T: what is it called?
L: sako
T: and?
L: aylana.

(e) Fillers
T: what is this object?
L: a sako um... sako wakaki.
T: great job.

Table 1: Dialogue Examples in the Data (L for the learner and T for the tutor)

Natural, spontaneous dialogue is *inherently incremental* (Crocker et al., 2000; Ferreira, 1996; Purver et al., 2009), and thus gives rise to dialogue phenomena such as self- and other-corrections, continuations, unfinished sentences, interruptions and overlaps, hedges, pauses and fillers. These phenomena are interactionally and semantically consequential, and contribute directly to how dialogue partners coordinate their actions and the emergent semantic content of their conversation. They also strongly mediate how a conversational agent might adapt to their partner over time. For example, self-interruption, and subsequent self-correction (see example in table 1.b) as well as hesitations/fillers (see example in table 1.e) aren't

simply noise and are used by listeners to guide linguistic processing (Clark and Fox Tree, 2002); similarly, while simultaneous speech is the bane of dialogue system designers, interruptions and subsequent continuations (see examples in table 1.c and 1.d) are performed deliberately by speakers to demonstrate strong levels of understanding (Clark, 1996).

Despite this importance, these phenomena are excluded in many dialogue corpora, and glossed over/removed by state of the art speech recognisers (e.g. Sphinx-4 (Walker et al., 2004) and Google's web-based ASR (Schalkwyk et al., 2010); see Baumann et al. (2016) for a comparison). One reason for this is that naturalistic spoken interaction is excessively expensive and time-consuming to transcribe and annotate on a level of granularity fine-grained enough to reflect the strict time-linear nature of these phenomena.

In this paper, we present a new dialogue data set - the BURCHAK corpus - collected using a new *incremental* variant of the DiET chat-tool (Healey et al., 2003; Mills and Healey, submitted)[1], which enables character-by-character, text-based interaction between pairs of participants, and which circumvents all transcription effort as all this data, including all timing information at the character level is automatically recorded.

The chat-tool is designed to support, elicit, and record at a fine-grained level, dialogues that resemble the face-to-face setting in that turns are: (1) constructed and displayed incrementally as they are typed; (2) transient; (3) potentially overlapping as participants can type at the same time; (4) not editable, i.e. deletion is not permitted - see Sec. 3 and Fig. 2. Thus, we have been able to collect many of the important phenomena mentioned above that arise from the inherently incremental nature of language processing in dialogue - see table 1.

Having presented the data set, we then go on to introduce a generic n-gram framework for building user simulations for either task-oriented or non-task-oriented dialogue systems from this dataset, or others constructed using the same tool. We apply this framework to train a robust user model that is able to simulate the tutor's behaviour to interactively teach (visual) word meanings to a Reinforcement Learning dialogue agent.

---

[1]Available from `https://sites.google.com/site/hwinteractionlab/babble`

(a) Chat Client Window (Tutor: "it is a ...", Learner: "suzuli?", Tutor: "sako burch")

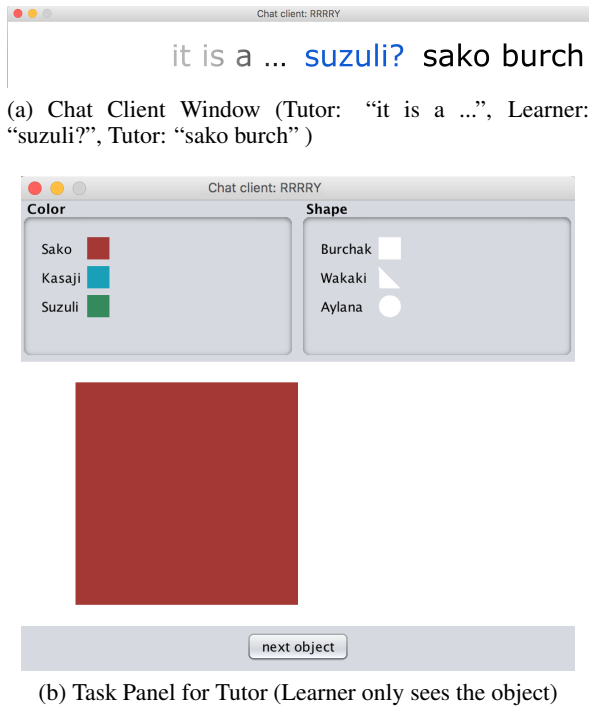(b) Task Panel for Tutor (Learner only sees the object)

Figure 2: Snapshot of the DiET Chat tool, the Tutor's Interface

2 Related Work

In this section, we will present an overview of relevant data-sets and techniques for Human-Human dialogue collection, as well as approaches to user simulation based on realistic data.

2.1 Human-Human Data Collection

There are several existing corpora of human-human spontaneous spoken dialogue, such as SWITCHBOARD (Godfrey et al., 1992), and the British National Corpus, which consist of open, unrestricted telephone conversations between people, where there are no specific tasks to be achieved. These datasets contain many of the incremental dialogue phenomena that we are interested in, but there is no shared visual scene between participants, meaning we cannot use such data to explore learning of perceptually grounded language. More relevant is the MAPTASK corpus (Thompson et al., 1993), where dialogue participants both have maps which are not shared. This dataset allows investigation of negotiation dialogue, where object names can be agreed, and so does support some work on language grounding. However, in the MAPTASK, grounded word meanings are not taught by ostensive definition as is the case in our new dataset.

We further note that the DiET Chat Tool (Healey et al., 2003; Mills and Healey, submitted) while designed to elicit conversational structures which resemble face-to-face dialogue (see examples in table 1), circumvents the need for the very expensive and time-consuming step of spoken dialogue transcription, but nevertheless produces data at a very fine-grained level. It also includes tools for creating more abstract (e.g. turn-based) representations of conversation.

2.2 User Simulation

Training a dialogue strategy is one of the fundamental tasks of the user simulation. Approaches to user simulation can be categorised based on the level of abstraction at which the dialogue is modeled: 1) the intention-level has become the most popular user model that predicts the next possible user dialogue action according to the dialogue history and the user/task goal (Eckert et al., 1997; Asri et al., 2016; Cuayáhuitl et al., 2005; Chandramohan et al., 2012; Eshky et al., 2012; Ai and Weng, 2008; Georgila et al., 2005); 2) on the word/utterance-level, instead of dialogue action, the user simulation can also be built for predicting the full user utterances or a sequence of words given specific information (Chung, 2004; Schatzmann et al., 2007b); and 3) on the semantic-level, the whole dialogue can be modeled as a sequence of user behaviors in the semantic representation (Schatzmann et al., 2007a; Schatzmann et al., 2007c; Kalatzis et al., 2016).

There are also some user simulations built on multiple levels. For instance, Jung et al. (2009) integrated different data-driven approaches on intention and word levels to build a novel user simulation. The user intent simulation is for generating user intention patterns, and then a two-phase data-driven domain-specific user utterance simulation is proposed to produce a set of structured utterances with sequences of words given a user intent and select the best one using the BLEU score. The user simulation framework we present below is generic in that one can use it to train user simulations on a word-by-word, utterance-by-utterance, or action-by-action levels, and it can be used for both goal-oriented and non-goal-oriented domains.

3 Data Collection using the DiET Chat Tool and a Novel Shape and Colour Learning Task

In this section, we describe our data collection method and process, including the concept learning task given to the human participants.

The DiET experimental toolkit This is a custom-built Java application (Healey et al., 2003; Mills and Healey, submitted) that allows two or more participants to communicate in a shared chat window. It supports live, fine-grained and highly local experimental manipulations of ongoing human-human conversation (see e.g. (Eshghi and Healey, 2015)). The variant we use here supports text-based, character-by-character, interaction between pairs of participants, and here we use it solely for data-collection, where everything that the participants type to each other passes through the DiET server, which transmits the utterance to the other clients on the character level and all are displayed *on the same row/track* in the chat window (see Fig. 2a) - this means that when participants type at the same time in interruptions and turn overlaps, their utterances will be all jumbled up (see Fig. 1b). To simulate the transience of speech in face-to-face conversation with its characteristic phenomena, all utterances in the chat window fade out after 1 second. Furthermore, like in speech, deletes are not permitted: if a character is typed, it cannot be deleted. The chat-tool is thus designed to support, elicit, and record at a fine-grained level, dialogues that resemble face-to-face dialogue in that turns are: (1) constructed and displayed incrementally as they are typed; (2) transient; (3) potentially overlapping; (4) not editable, i.e. deletion is not permitted.

Task and materials The learning/tutoring task given to the participants involves a pair of participants who talk about visual attributes (e.g. colour and shape) through a sequence of 9 visual objects, one at a time. The objects are created based on a 3 x 3 visual attribute matrix (including 3 colours and 3 shapes (see Fig.2b)). This task is assumed in a second-language learning scenario, where each visual attribute, instead of standard English words, is assigned to a new unknown word in a made-up language, e.g. "sako" for red and "burchak" for square: participants are not allowed to use any of the usual colour and shape words from the English language. We design the task in this way to col-lect data for situations where a robot has to learn the meaning of human visual attribute terms. In such a setting the robot has to learn the perceptual groundings of words such as "red". However, humans already know these groundings, so to collect data about teaching such perceptual meanings, we invented new attribute terms whose groundings the Learner must discover through interaction.

The overall goal of the task is for the learner to identify the shape and colour of the presented objects correctly for as many objects as possible. So the tutor initially needs to teach the learner about these using the presented objects. For this, the tutor is provided with a visual dictionary of the (invented) colour and shape terms (see Fig. 2), but the learner only ever sees the object itself. The learner will thus gradually learn these and be able to identify them, so that initiative in the conversation tends to be reversed on later objects, with the learner making guesses and the tutor either confirming these or correcting them.

Participants Forty participants were recruited from among students and research staff from various disciplines at Heriot-Watt University, including 22 native speakers and 18 non-native speakers.

Procedure The participants in each pair were randomly assigned to experimental roles (Tutor vs. Learner). They were given written instructions about the task and had an opportunity to ask questions about the procedure. They were then seated back-to-back in the same room, each at a desk with a PC displaying the appropriate task window and chat client window (see Fig.2). They were asked to go through all visual objects in at most 30 minutes and then the Learner was assessed to check how many new colour and shape words they had learned. Each participant was paid 10.00 for participation. The best performing pair was also given a 20 Amazon Voucher as prize.

4 The BURCHAK Corpus Statistics

4.1 Overview

Using the above procedure, we have collected 177 dialogues (each about one visual object) with a total of 2454 turns, where a turn is defined[2] as a sequence of consecutive characters typed by a single participant with a delay of no more than 1100 ms

---

[2] Note that the definition of a 'turn' in an incremental system is somewhat arbitrary.

between the characters. Figure 4a shows the distribution of dialogue length (i.e. number of turns) in the corpus. where the average number of turns per dialogue is 13.86.

4.2 Incremental Dialogue Phenomena

As noted, the DiET Chattool is designed to elicit and record conversations that resemble face-to-face dialogue. In this paper, we report specifically on a variety of dialogue phenomena that arise from the incremental nature of language processing. These are the following:

- **Overlapping**: where interlocutors speak/type at the same time (i.e. the original corpus contains over 800 overlaps), leading to jumbled up text on the DiET interface (see Fig. 1);

- **Self-Correction**: a kind of correction that is performed incrementally in the same turn by a speaker; this can either be conceptual, or simply repairing a misspelling or mispronunciation.

- **Self-Repetition**: the interlocutor repeats words, phrases, even sentences, in the same turn.

- **Continuation (aka Split-Utterance)**: the interlocutor continues the previous utterance (by herself or the other) where either the second part, or the first part or both are syntactically incomplete.

- **Filler**: allows the interlocutor to further plan her utterance while keeping the floor. These can also elicit continuations from the other (Howes et al., 2012). This is performed using tokens such as 'urm', 'err', 'uhh', or '...'.

For annotating self-corrections, self-repetitions and continuations we have loosely followed protocols from Purver et al. (2009; Colman and Healey (2011). Figure 4d shows how frequently these incremental phenomena occur in the BURCHAK Corpus. This figure excludes Overlaps which were much more frequent: 800 in total, which amounts to about 4.5 per dialogue.

4.3 Cleaning up the data for the User Simulation

For the purpose of the annotation of Dialogue Actions, subsequent training of the user simulation, and the Reinforcement Learning described below, we cleaned up the original corpus as follows: 1)

we fixed the spelling mistakes which were not repaired by the participants themselves; 2) we also removed snippets of conversation where the participants had misunderstood the task (e.g. trying to describe the objects or where they had used other languages) (see Figure 3); as well as 3) removing emoticons (which frequently occurs in the chat tool).

T: the word for the color is similar to the word for Japanese rice wine. except it ends in o.
L: sake?
T: yup, but end with an o.
L: okay, sako.

Figure 3: Example of Dialogue Snippet with the misunderstanding of the task

We trained a simulated tutor based on this cleaned up data (see below, Section 5).

4.4 Dialogue Actions and their frequencies

The cleaned up data was annotated for the following dialogue actions:

- **Inform**: the action to inform the correct attribute words of an object to the partner, including statement, question-answering, correction, , e.g. "this is a suzuli burchak" or "this is sako";

- **Acknowledgment**: the ability to process confirmations from the *tutor*/the *learner*, e.g. "Yes, it's a square".

- **Rejection**: the ability to process negations from the *tutor*, e.g. "no, it's not red";

- **Asking**: the action to ask WH or polar questions requesting correct information, e.g. "what colour is this?" or "is this a red square?".

- **Focus**: the action to switch the dialogue topic onto specific objects or attributes, e.g. "let's move to shape now";

- **Clarification**: the action to clarify the categories for particular attribute names, e.g. "this is for color not shape";

- **Checking**: the action to check whether the partner understood, e.g. "get it?";

- **Repetition**: the action to request Repetitions to double-check the learned knowledge, e.g. "can you repeat the color again?";

- **Offer-Help**: the action to help the partner answer questions, occurs frequently when the learner cannot answer it immediately, e.g. "L: it is a ... T: need help? L: yes. T: a sako burchak.";

Fig. 4c shows how often each dialogue action occurs in the data set; and Fig. 4b shows the frequencies of these actions by the learner and the tutor individually in each dialogue turn. In contrast with a lot of previous work which assumes a single action per turn, here we get multiple actions per turn (see Table 1) In terms of the *Learner* behavior, the learner mostly performs a single action per turn. On the other hand, although the majority of the dialogue turns on the tutor side also have a single action, about 22.59% of the dialogue turns perform more than one action.

5 TeachBot User Simulation

Here we describe the generic user simulation framework, based on n-grams, for building user simulation from this type of incremental corpus. We apply this framework to train a TeachBot user simulator that is used to train a RL interactive concept learning agent, both here, and in future work. The model is here trained from the cleaned up version of the corpus.

5.1 The N-gram User Simulation

The proposed user model is a compound n-gram simulation that the probability $(P(t|w_1, .., w_n, c_1, .., c_m))$ of an item t (an action or utterance from the tutor in our work) is predicted based on a sequence of the most recent words $(w_1, \ldots, w_n)$ from the previous utterance and additional dialogue context parameters C:

$$P(t|w_1, .., w_n, c_1, .., c_m) = \frac{freq(t, w_1, .., w_n, c_1, .., c_m)}{freq(w_1, .., w_n, c_1, .., c_m)}$$
(1)

where $c_1, .., c_m \in C$ represent additional conditions for specific user/task goals (e.g. goal completion as well as previous dialogue context).

For this specific task, the additional dialogue conditions (C) are as follows: (1) the color state (C_{state}) for whether the color attribute is identified correctly, (2) the shape state (S_{state}) for whether the shape attribute is identified correctly, as well as 3) the previous context ($preContxt$) for which attribute (colour or shape) is currently under discussion.

In order to reduce mismatch risk, the simulation model is able to back-off to smaller n-grams when it cannot find any n-grams matched to the current word sequence and conditions. To eliminate the search restriction by the additional conditions, we applied the nearest neighbors algorithm to search for the n-gram matches by calculating the Hamming distance of each pair of n-grams.

The n-gram user simulation is generic, as it is designed to handle the item prediction on multiple levels, on which the predicted item, t, can be assigned either to (1) a full user utterance (U_t) on the utterance level; (2) a combined sequence of dialogue actions (Das_t); or alternatively (3) the next word/lexical token. During the simulation, the n-gram model chooses the next item according to the distribution of n-grams. In terms of the action level, a user utterance will be chosen upon a distribution of utterance templates collected from the corpus and combined given dialogue actions Das_t. The tutor simulation we train here is at the level of the action and utterance, and is evaluated on the same levels below. However, the framework can be used to train to predict fully incrementally on a word-by-word basis. In this case, the $w_i (i < n)$ in Eq.1 will contain not only a sequence of words from the previous system utterance, but also words from the current speaker (the tutor itself as it is generating).

The probability distribution in equation 1 is induced from the corpus using Maximum Likelihood Estimation, where we count how many times each t occurs with any specific combination of the conditions $(w_1, \ldots, w_n, c_1, \ldots, c_m)$ and divide this by the total number of times t occurs (see Eq 1).

5.2 Evaluation of the User Simulation

We evaluate the proposed user simulation based on the turn-level evaluation metrics by (Keizer et al., 2012), in which evaluation is done on a turn-by-turn basis. Evaluation is done based on the cleaned up corpus (see Section 4). We investigate the performance of the user model on two levels: the utterance level and the action level.

The evaluation is done by comparing the distribution of the predicted actions or utterances with the actual distributions in the data. We report two measures: the Accuracy and Kullback-Leibler Divergence (cross-entropy) to quantify how closely the simulated user responses resemble the real user
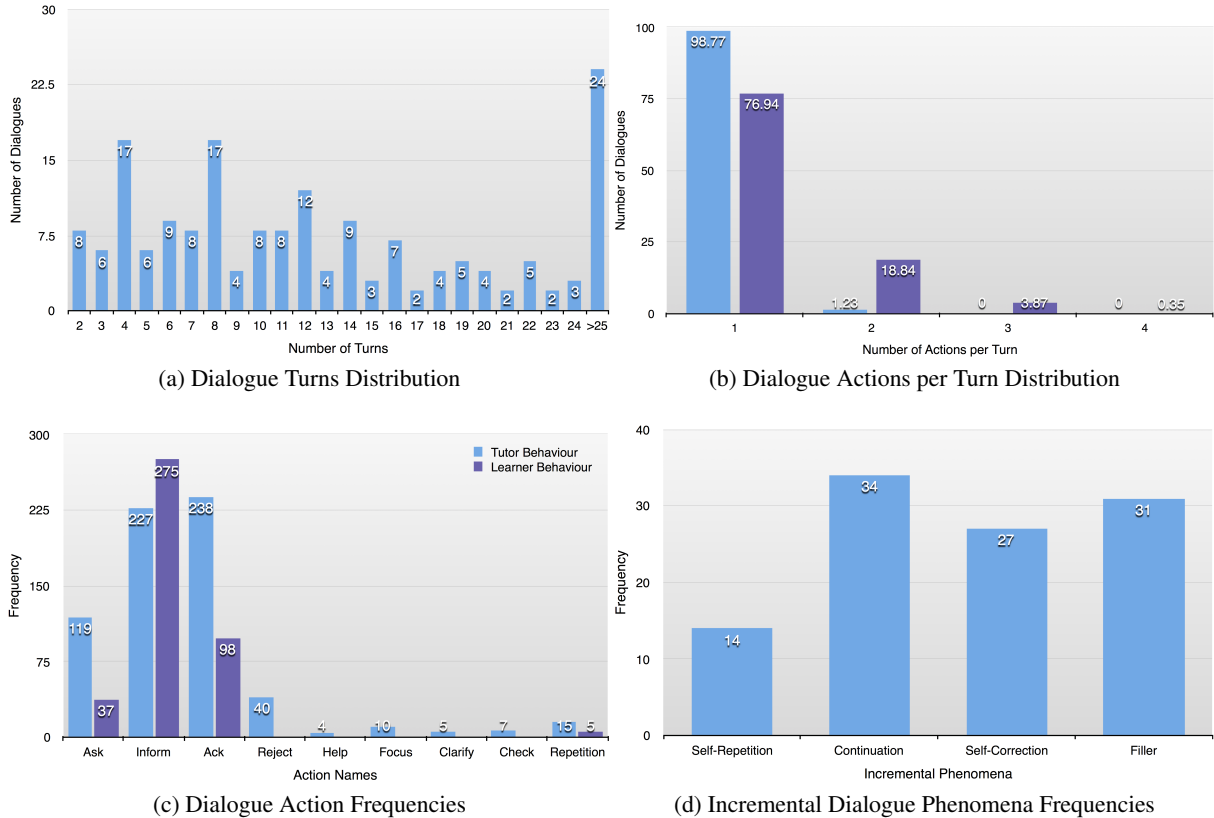
(a) Dialogue Turns Distribution

(b) Dialogue Actions per Turn Distribution

(c) Dialogue Action Frequencies

(d) Incremental Dialogue Phenomena Frequencies

Figure 4: Corpus Statistics

responses in the BURCHAK corpus. Accuracy (Acc) measures the proportion of times an utterance or dialogue act sequence (Das_t) is predicted correctly by the simulator, given a particular set of conditions ($w_1, .., w_n, c_1, .., c_m$). To calculate this, all existing combinations in the data of the values of these variables are tried. If the predicted action or utterance occurs in the data for these given conditions, we count the prediction as correct.

Kullback-Leibler Divergence (KLD) ($D_{kl}(P \parallel Q)$) is applied to compare the predicted distributions and the actual one in the corpus (see Eq.2).

$$D_{kl}(P \parallel Q) = \sum_{i=1}^{M} p_i \log(\frac{p_i}{q_i}) \qquad (2)$$

Table 2 shows the results: the user simulation on both utterance and action levels achieves good performance. The action-based user model, on a more abstract level, would likely be better as it is less sparse, and produces more variation in the resulting utterances.

Ongoing work involves using BURCHAK to train a word-by-word incremental tutor simulation, capable of generating all the incremental phenomena identified earlier.

Simulation	Accuracy (%)	KLD
Utterance-level	77.98	0.2338
Act-level	**84.96**	**0.188**

Table 2: Evaluation of The User Simulation on both Utterance and Act levels

6 Training a prototype concept learning agent from the BURCHAK corpus

In order to demonstrate how the BURCHAK corpus can be used, we train and evaluate a prototype interactive learning agent using Reinforcement Learning (RL) on the collected data. We follow previous task and experiment settings (see (anon, anon)) to compare the learned RL-based agent with a rule-based agent with the best performance from previous work. Instead of using hand-crafted dialogue examples as before, here we train the RL agent in interaction with the user simulation, itself trained from the BURCHAK data as above.

6.1 Experiment Setup

To compare the performance of the rule-based system and the trained RL-based system in the interactive learning process, we follow all experi-

ment setup, including visual data-set and cross-validation method. We also follow the evaluation metrics provided by (2016b) : *Overall Performance Ratio* (R_{perf}) to measures the trade-offs between the cost to the tutor and the accuracy of the learned meanings, i.e. the classifiers that ground our colour and shape concepts. (see Eq.3).

$$R_{perf} = \frac{\Delta Acc}{C_{tutor}} \quad (3)$$

i.e. the increase in accuracy per unit of the cost, or equivalently the gradient of the curve in Fig. 5 We seek dialogue strategies that maximise this.

The cost C_{tutor} measure reflects the effort needed by a human tutor in interacting with the system. Skocaj et. al. (2009) point out that a comprehensive 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: C_{inf} refers to the cost (assigned to 5 points) of the tutor providing information on a single attribute concept (e.g. "this is red" or "this is a square"); $C_{ack/rej}$ is the cost (0.5 points) for a simple confirmation (like "yes", "right") or rejection (such as "no"); C_{crt} is the cost of correction (5 points) for a single concept (e.g. "no, it is blue" or "no, it is a circle").

6.2 Results & Discussion

Fig. 5 plots Accuracy against Tutoring Cost directly. The gradient of this curve corresponds to *increase in Accuracy per unit of the Tutoring Cost*: a measure of the trade-off between accuracy of learned meanings and tutoring cost.

The result shows that the RL-based learning agent achieves a comparable performance with the rule-based system.
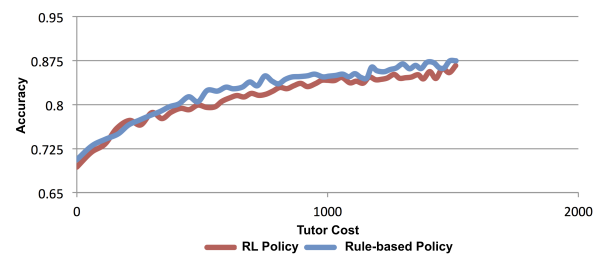

Figure 5: Evolution of Learning Performance

Table 3 shows an example dialogue between the learned concept learning agent and the tutor simulation, where the user model simulates the tutor behaviour (*T*) for the learning tasks. In this example, the utterance produced by the simulation involves two incremental phenomena, i.e. a self-correction and a continuation, though note that these have not been produced on a word-by-word level.

L: so is this shape square?
T: no, it's a squ ... sorry ... a circle. and color?
L: red?
T: yes, good job.

Table 3: Dialogue Example between a Learned Policy and the Simulated Tutor

7 Conclusion

We presented a new data collection tool, a new data set, and and associated dialogue simulation framework which focuses on visual language grounding and natural, incremental dialogue phenomena. The tools and data are freely available and easy to use.

We have collected new human-human dialogue data on visual attribute learning tasks, which are then used to create a generic n-gram user simulation for future research and development. We used this n-gram user model to train and evaluate an optimized dialogue policy, which learns grounded word meanings from a human tutor, incrementally, over time. This dialogue policy optimisation learns a complete dialogue control policy from the data, in contrast to earlier work (Yu et al., 2016b) which only optimised confidence thresholds, and where dialogue control was entirely rule-based.

Ongoing work further uses the data and simulation framework here to train a word-by-word incremental tutor simulation, with which to learn complete, incremental dialogue policies, i.e. policies that choose system output at the lexical level (Eshghi and Lemon, 2014). To deal with uncertainty this system in addition takes all the visual classifiers' confidence levels directly as features in a continuous space MDP.

References

Hua Ai and Fuliang Weng. 2008. User simulation as testing for spoken dialog systems. In *Proceedings of the SIGDIAL 2008 Workshop, The 9th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 19-20 June 2008, Ohio State University, Columbus, Ohio, USA*, pages 164–171.

Layla El Asri, Jing He, and Kaheer Suleman. 2016. A sequence-to-sequence model for user simulation in spoken dialogue systems. *CoRR*, abs/1607.00070.

Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2016. Recognising conversational speech: What an incremental asr should do for a dialogue system and how to get there. In *International Workshop on Dialogue Systems Technology (IWSDS) 2016*. Universität Hamburg.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(1–47).

Senthilkumar Chandramohan, Matthieu Geist, Fabrice Lefevre, and Olivier Pietquin. 2012. Behavior specific user simulation in spoken dialogue systems. In *Speech Communication; 10. ITG Symposium; Proceedings of*, pages 1–4. VDE.

Grace Chung. 2004. Developing a flexible spoken dialog system using simulation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain.*, pages 63–70.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84(1):73–111.

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

M. Colman and P. G. T. Healey. 2011. The distribution of repair in dialogue. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 1563–1568, Boston, MA.

Matthew Crocker, Martin Pickering, and Charles Clifton, editors. 2000. *Architectures and Mechanisms in Sentence Comprehension*. Cambridge University Press.

Heriberto Cuayáhuitl, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. 2005. Human-computer dialogue simulation using hidden markov models. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 290–295. IEEE.

Wieland Eckert, Esther Levin, and Roberto Pieraccini. 1997. User modeling for spoken dialogue system evaluation. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 80–87. IEEE.

Arash Eshghi and Patrick G. T. Healey. 2015. Collective contexts in conversation: Grounding by proxy. *Cognitive Science*, pages 1–26.

Arash Eshghi and Oliver Lemon. 2014. How domain-general can we be? learning incremental dialogue systems without dialogue acts. In *Proceedings of SemDial*.

Aciel Eshky, Ben Allison, and Mark Steedman. 2012. Generative goal-driven user simulation for dialog management. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 71–81.

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*.

Victor Ferreira. 1996. Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language*, 35:724–755.

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2005. Learning user simulations for information state update dialogue systems. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 893–896.

John J. Godfrey, Edward Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of IEEE ICASSP-92*, pages 517–520, San Francisco, CA.

P. G. T. Healey, Matthew Purver, James King, Jonathan Ginzburg, and Greg Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Boston, Massachusetts, August.

Christine Howes, Patrick G. T. Healey, Matthew Purver, and Arash Eshghi. 2012. Finishing each other's ... responding to incomplete contributions in dialogue. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci 2012)*, pages 479–484, Sapporo, Japan, August.

Sangkeun Jung, Cheongjae Lee, Kyungduk Kim, Minwoo Jeong, and Gary Geunbae Lee. 2009. Data-driven user simulation for automated evaluation of spoken dialog systems. *Computer Speech & Language*, 23(4):479–509.

Dimitrios Kalatzis, Arash Eshghi, and Oliver Lemon. 2016. Bootstrapping incremental dialogue systems: using linguistic knowledge to learn from minimal data. In *Proceedings of the NIPS 2016 workshop on Learning Methods for Dialogue*, Barcelona.

Simon Keizer, Stphane Rossignol, Senthilkumar Chandramohan, and Olivier Pietquin. 2012. User Simulation in the Development of Statistical Spoken Dialogue Systems. In Oliver Lemon Olivier Pietquin, editor, *Data-Driven Methods for Adaptive Spoken Dialogue Systems: Computational Learning for Conversational Interfaces*, chapter 4, pages 39–73. Springer, November.

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.

Gregory J. Mills and Patrick G. T. Healey. submitted. The Dialogue Experimentation toolkit. *xx*, (?).

Matthew Purver, Christine Howes, Eleni Gregoromichelaki, and Patrick G. T. Healey. 2009. Split utterances in dialogue: A corpus study. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009 Conference)*, pages 262–271, London, UK, September. Association for Computational Linguistics.

Johan Schalkwyk, Doug Beeferman, Franoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. 2010. "Your word is my command": Google search by voice: A case study. In *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, chapter 4, pages 61–90. Springer, New York.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve J. Young. 2007a. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA*, pages 149–152.

Jost Schatzmann, Blaise Thomson, and Steve J. Young. 2007b. Error simulation for training statistical dialogue systems. In *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2007, Kyoto, Japan, December 9-13, 2007*, pages 526–531.

Jost Schatzmann, Blaise Thomson, and Steve J. Young. 2007c. Statistical user simulation with a hidden agenda. In *Proceedings of the SIGDIAL 2007 Workshop, The 9th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 1-2 September 2007, the University of Antwerp, Belgium*.

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, 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.

Henry S. Thompson, Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The hcrc map task corpus: Natural dialogue for speech recognition. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 25–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. 2004. Sphinx-4: A flexible open source framework for speech recognition. Technical report, Mountain View, CA, USA.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016a. Incremental generation of visually grounded language in situated dialogue. In *Proceedings of INLG 2016*.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016b. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, pages 339–349.

10