# Incrementally Tracking Reference in Human/Human Dialogue Using Linguistic and Extra-Linguistic Information

**Casey Kennington**
CITEC, Bielefeld University
Universitätsstraße 25
Bielefeld Germany
ckennington@cit-ec.uni-bielefeld.de

**Ryu Iida**
NICT UCRI
Seika-cho, Soraku-gun
Kyoto, Japan
ryu.iida@nict.go.jp

**Takenobu Tokunaga**
Tokyo Institute of Technology
Ôokayama, Meguro
Tokyo, Japan
take@cs.titech.ac.jp

**David Schlangen**
Bielefeld University
Universitätsstraße 25
Bielefeld Germany
david.schlangen@uni-bielefeld.de

## Abstract

A large part of human communication involves referring to entities in the world and often these entities are objects that are visually present for the interlocutors. A system that aims to resolve such references needs to tackle a complex task: objects and their visual features need to be determined, the referring expressions must be recognised, and extra-linguistic information such as eye gaze or pointing gestures need to be incorporated. Systems that can make use of such information sources exist, but have so far only been tested under very constrained settings, such as WOz interactions. In this paper, we apply to a more complex domain a reference resolution model that works incrementally (i.e., word by word), grounds words with visually present properties of objects (such as shape and size), and can incorporate extra-linguistic information. We find that the model works well compared to previous work on the same data, despite using fewer features. We conclude that the model shows potential for use in a real-time interactive dialogue system.

## 1 Introduction

Referring to entities in the world via definite descriptions makes up a large part of human communication (Poesio and Vieira, 1997). In task-oriented situations, these references are often to entities that are visible in the shared environment. This kind of reference has attracted attention in recent computational research, but the kinds of interactions studied are often fairly restricted in controlled lab situations (Tanenhaus and Spivey-Knowlton, 1995) or simulated human/computer interactions, (Schlangen et al., 2009; Kousidis et al., 2013; Chai et al., 2014). In such task-oriented, co-located settings, interlocutors can make use of extra-linguistic cues such as gaze or pointing gestures. Furthermore, listeners resolve references as they unfold, often identifying the referred entity before the end of the reference (Tanenhaus and Spivey-Knowlton, 1995; Spivey et al., 2002), however research in reference resolution has mostly focused on full, completed referring expressions.

In this paper we make a first move towards addressing somewhat more complex domains. We apply a model of reference resolution, which has been tested in a simpler setup, on more natural data coming from a corpus of human/human interactions. The model is *incremental* in that it does not wait until the end of an utterance to process, rather it updates its interpretation at each word increment. The model can also incorporate other modalities, such as gaze or pointing cues (deixis) incrementally. We also model the saliency of the context, and show that the model can easily take such contextual information into account. The model improves over previous work on reference resolution applied to the same data (Iida et al., 2010; Iida et al., 2011).

The paper is structured as follows: in the following section we discuss related work on incremental resolution of referring expressions. We explain the model that we use in Section 3 and the data we apply it to in Section 4. We then describe the experiments and the results and provide a discussion.

## 2 Related Work

*Reference resolution* (RR), which is the task of resolving referring expressions (REs) to what they are



*Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 272–282,
Denver, Colorado, May 31 – June 5, 2015. 

intended to refer to, has been well-studied in various fields such as psychology (Isaacs and Clark, 1987; Tanenhaus and Spivey-Knowlton, 1995), linguistics (Pineda and Garza, 2000), as well as human/human (Iida et al., 2010) and human/machine interaction (Prasov and Chai, 2010; Siebert and Schlangen, 2008; Schlangen et al., 2009). In recent years, multi-modal corpora have emerged which provide RR with important contextual information: collecting dialogue between two humans (Tokunaga et al., 2012; Spanger et al., 2012), between a human and a (simulated) dialogue system (Kousidis et al., 2013; Liu et al., 2013), with gaze, information about the shared environment, and in some cases deixis.

It has been shown that incorporating gaze improves RR in a situated setting because speakers need to look at and distinguish from distractors the objects they are describing: this has been shown in a static scene on a computer screen (Prasov and Chai, 2008), in human-human interactive puzzle tasks (Iida et al., 2010; Iida et al., 2011), in web browsing (Hakkani-tür et al., 2014), and in a moving car where speakers look at objects in their vicinity (Misu et al., 2014). Incorporating pointing (deictic) gestures is also potentially useful in situated RR; as for example Matuszek et al. (2014) have shown in work on resolving objects processed by computer vision techniques. Chen and Eugenio (2012) looked into reference in multi-modal settings, with focus on co-referential pronouns and pointing gestures. However, these approaches were applied in settings in which communication between the two interlocutors was constrained, or the developed systems did not process incrementally. Kehler (2000) presented approach that focused more on interaction in a map task, though the model was not incremental, nor did grounding occur between language and world, as we do here.

Incremental RR has also been studied in a number of papers, including a framework for fast incremental interpretation (Schuler et al., 2009), a Bayesian filtering model approach that was sensitive to disfluencies (Schlangen et al., 2009), a model that used Markov Logic Networks to resolve objects on a screen (Kennington and Schlangen, 2013), a model of RR and incremental feedback (Traum et al., 2012), and an approach that used a semantic representation to refer to objects (Peldszus et al., 2012; Kennington et al., 2014). However, the approaches reported there did not incorporate multi-modal information, were too slow to work in real-time, were evaluated on constrained data, or only focused on a specific type of RR, ignoring pronouns or deixis.

In this paper, we opted to use the model presented in Kennington et al. (2013), the *simple incremental update model* (SIUM). It has been tested extensively against data from a puzzle-playing human/computer interaction domain (the PENTO data, (Kousidis et al., 2013)); it can incorporate multi-modal information, works in real-time, and can resolve definite, exophoric, and deictic references in a single framework, all of which makes it a potential candidate for working in an interactive, multi-modal dialogue system. The model is similar to the one proposed in Funakoshi et al. (2012), which could resolve descriptions, anaphora, and deixis in a unified manner, but that model does not work incrementally.[1]

The main contributions of this paper are the more thorough exposition of the model (in Section 3) and its application and evaluation on much less constrained, more interactive (and hence realistic) data than what it has previously been tested on (Section 4). Moreover, the data set used here is also from a typologically very different language (Japanese) than what the model has been previously tested on (German), and so the robustness of the model against these differences is also investigated.

We will now describe the model, and that will be followed by a description of the corpus we used.

## 3 The Simple Incremental Update Model

Following Kennington et al. (2013) and Kennington et al. (2014), we model the task at hand as one of recovering $I$, the intention of the speaker making the RE, where $I$ ranges over the possible alternatives (the objects in the domain). This recovery proceeds incrementally (word by word), for RE of arbitrary length. That is, if $U$ denotes the current word, we are interested in $P(I|U)$, the current hypothesis about

[1] It can be argued that any non-incremental model could be made into an incremental one by applying that model at each word (Khouzaimi et al., 2014), but we would argue that more modeling effort is required in order for the model to work in an interactive dialogue system, see (Schlangen and Skantze, 2009; Aist et al., 2007; Skantze and Schlangen, 2009; Skantze and Hjalmarsson, 1991).

the intended referent, given the observed word. We assume the presence of an unobserved, latent variable $R$, which models properties of the candidate objects such as colour or shape; explained further below), and so the computation formally is:

$$P(I|U) = \sum_{r \in R} \frac{P(I, U, R)}{P(U)} \quad (1)$$

Which, after making some independence assumptions, can be factored into:

$$P(I|U) = \frac{1}{P(U)} P(I) \sum_{r \in R} P(U|R) P(R|I) \quad (2)$$

This is an update model in the usual sense that the posterior $P(I|U)$ at one step becomes the prior $P(I)$ at the next. $P(R|I)$ provides the link between the intentions (that is, objects) and the properties (e.g., the colour and shape of each object), and $P(U|R)$ the link between properties and (observed) words. Being incremental, this model is computed at each word. As properties play an important role in this model, they will now be explained.

**Properties** The variable $R$ models visual or abstract properties of entities (such as real-world objects or linguistic entities) and their selection for verbalisation in the referring expression. The simple assumption made by the model is that only such properties can be selected for verbalisation which the candidate object actually has. Hence, the starting point for the model is a representation of the world and the current dialogue context in terms of the properties of the objects. For this paper, this means properties belonging to objects in the shared work space.

We will explain the properties we used in our implementation of this model (henceforth SIUM, i.e., *simple incremental update model*), the motivation for using them, and give an example of applying the model in Section 5.

## 4 The REX Data

The corpora presented in Iida et al. (2011) and Spanger et al. (2012) are a collection of human/human interaction data where the participants collaboratively solved Tangram puzzles. In this setting, anaphoric references (i.e., pronoun references to entities in an earlier utterance, e.g., "move *it* to the left") and exophoric references via definite descriptions (i.e., references to real-world objects, e.g., "*that* one" or "the big triangle") are common (note that both refer in different ways to objects that are physically present). The corpus also records an added modality: the gaze of the puzzle *solver* (SV) who gives the instructions and that of the *operator* (OP), who moves the tangram pieces. The mouse pointer controlled by the OP could also be considered a modality, used as a kind of pointing gesture that both participants can observe. The goal of the task was to arrange puzzle pieces on a board into a specified shape (example in Figure 1), which was only known to SV and hidden from OP. The language of the dialogues was Japanese.

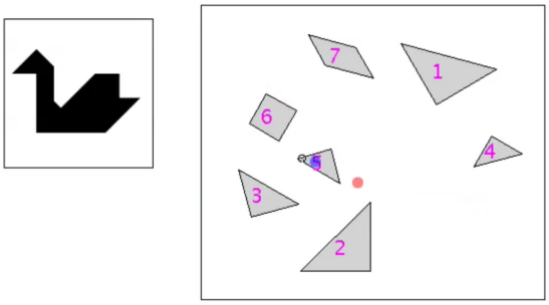


Figure 1: Example Tangram Board; the goal shape is the swan in the top left, the shared work area is the large board on the right, the mouse cursor and OP gaze (blue dot) are on object 5, the SV gaze is the red dot (gaze points were not seen by the participants).

This environment provided frequent use of REs that aimed to distinguish puzzle pieces (and piece groups) from each other. The following are some example REs from the REX corpus:

(1) a. *chicchai sankakkei*
b. small triangle

(2) a. *sono ichiban migi ni shippo ni natte iru sankakkei*
b. that most right tail becoming triangle
'that right-most triangle that is the tail'

Example (1) is a typical example of an RE as found in the corpus. Note that this at the same time constitutes the whole utterance, which hence can be classi-

fied as a non-sentential utterance (Schlangen, 2004). Its transliteration consists of 8 Japanese characters, which could be tokenized into two words. The more difficult RE shown in Example (2) requires the model to learn how spatial placements map to certain descriptions. Moreover, Japanese is a head-final language where comparative landmark pieces are uttered before the referent. Also, because this was a highly interactive setting, many exophoric pronouns were used, e.g., *sore* and *sono*, both meaning *that*.[2] Pronoun references like this made up around 32% of the utterances.

Corpus annotations included (for both participants) transcriptions of utterances, the object being looked at any given time, the object being pointed at or manipulated by the mouse, segmentation of the REs and the corresponding referred object or objects. The spatial layout of the board was recorded each time an object was manipulated. Further details of the corpus can be found in (Iida et al., 2011). In order to directly compare our work with previous work, in our evaluations below we consider the same annotated REs. Iida et al. (2011) applied a support vector machine-based ranking algorithm (Joachims, 2002) to the task of resolving REs in this corpus. They used a total of 36 binary features in the SVM classifier, which predicted the referred object. They further used a separate model for pronoun utterances and non-pronoun utterances, allowing the classifier to learn patterns without confusing utterance types. More details on the results of these models are given below.

The SIU-model has previously been applied to two datasets from the Pentomino domain (Kennington et al., 2013), where the speaker's goal was to identify one out of a set of tetris-like (but consisting of five instead of four blocks) puzzle pieces. However, in these datasets, the references were "one-shot" and not embedded in longer dialogues, as is the case in the REX corpus. A summary of differences between the two tasks is summarised in Table 1. Applying SIUM to data like that found in the REX corpus is a natural next step to test the abilities of the model as a RR component in a dialogue system.

[2]To be precise, *sono* is a demonstrative adjective.

| | PENTO | REX |
|---|---|---|
| language | **German** | **Japanese** |
| language type | SVO | SOV |
| phrase type | head-initial | head-final |
| avg utt length | 7-8 | 4-5 |
| number of objects | 15 | 7 |
| interactivity | human-wizard | human-human |
| recorded gaze | SV (speaker) | SV, OP |
| % of pronoun utts | 0% | 32% |

Table 1: Summary of differences between PENTO and REX tasks.

## 5 Experiment

**Procedure** The procedure for this experiment is as follows. In order to compare our results directly with those of Iida et al. (2011), we provide our model with the same training and evaluation data, in a 10-fold cross-validation of the 1192 REs from 27 dialogues (the T2009-11 corpus in Tokunaga et al. (2012)). For development, we used a separate part of the REX corpus (N2009-11) that was structured similarly to the one used in our evaluation.

**Task** The task is RR. At each increment, SIUM returns a distribution over all objects; the probability for each object represents the strength of the belief that it is the referred one. The argmax of the distribution is chosen as the hypothesised referred object.

**P(R|I)** $P(R|I)$ models the likelihood of selecting a property of a candidate object for verbalisation; this likelihood is assumed to be uniform for all the properties that the candidate object has.[3] We derive these properties from a representation of the scene; similar to how Iida et al. (2011) computed features to present to their classifier: namely **Ling** (linguistic features), **TaskSp** (task specific features), and **Gaze** (from SV only). Some features were binary, others such as shape and size had more values. Table 2 shows all the properties that were used here. Each will now be explained.

**Ling** Each object had a shape, size, and relative position to the other pieces. We determined by hand

[3]Uniformity in the likelihood of the properties isn't an ideal approach as certain properties could be more likely to be selected than others; we leave a more principled approach to using saliency to help determine the likelihood of the properties to future work.

| Ling | TaskSp |
|---|---|
| tri/squ/pgram | most_recent_move |
| small/med/big | mouse_pointed |
| left/mid/right | |
| prev_referred | **Gaze** |
| top/cen/bottom | most_gazed_at |
| referred_5 | gazed_at_in_utt |
| referred_10 | longest_gazed_at |
| referred_20 | recent_fixation |

Table 2: List of properties used for each source of information.

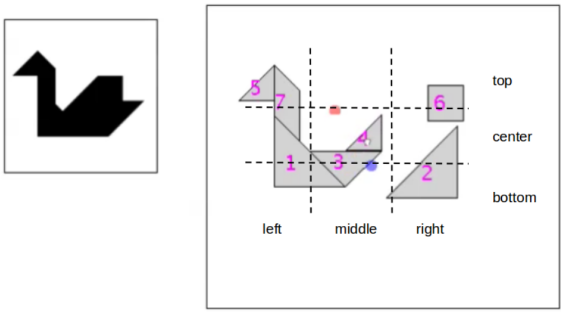


Figure 2: Tangram later in the dialogue; the notion of *right-ness* and other spatial concepts changes throughout the dialogue (compare to Figure 1), the grids are added to show which objects receive which horizontal and which vertical properties.

the shape and size properties which remained static through each dialogue. The position properties were derived from the corpus logs. For each object, the centroid of each object was computed. Then, the vertical and horizontal range for all of the objects was calculated and then split into three even sections in each dimension (see Figure 2). An object with a centroid in the left-most section of the horizontal range received a `left` property, similarly `middle` and `right` properties were calculated for corresponding objects. For vertical placement, `top`, `center` and `bottom` properties were given to objects in the respective vertical segments. Figure 2 shows an example segmentation. Each object had a vertical and a horizontal property at all times, however, moving an object could result in a change of one of these spatial properties as the dialogue progressed. As an example, compare Figure 1, which is a snapshot of the interaction towards the beginning, and Figure 2, which shows a later stage of the game board; spatial layout changes throughout the dialogue.

These properties differ somewhat from the features for the Ling model presented in Iida et al. (2011). Three features that we did use as properties had to do with reference recency: the most recently referred object received the `referred_X` properties, if an object was referred to in the past 5, 10, or 20 seconds.

**TaskSp** Iida et al. (2011) used 14 task-specific features, three of which they found to be the most informative in their model. Here, we will only use the two most informative features as properties (the third one, whether or not an object was being manipulated at the beginning of the RE, did not improve results in a held-out test): the object that was most recently moved received the `most_recent_move` property and objects that have the mouse cursor over them received the `mouse_pointed` property (see Figure 2; object 4 would receive both of these properties, but only for the duration that the mouse was actually over it). Each of these properties can be extracted directly from the corpus annotations.

**Gaze** Similar to Iida et al. (2011), we consider gaze during a window of 1500ms before the onset of the RE. The object that was gazed at the longest during that time received a `longest_gazed_at` property, the object which was fixated upon most recently during that interval before the RE onset received a `recent_fixation` property, and the object which had the most fixations received the `most_gazed_at` property. During a RE, an object received the `gazed_at_in_utt` property if it is gazed at during the RE up until that point. These properties can be extracted directly from the corpus annotations. Other gaze features are not really accessible to an incremental model such as this, as gaze features extracted from gaze activity over the RE can only be computed when it is complete. Our Gaze properties are made up of these 4 properties, as opposed to the 14 features in Iida et al. (2011).

**P(U|R)** $P(U|R)$ is the model that connects the property selected for verbalisation with a way of verbalising it (a value for $U$). Instead of directly learning this model from data, which would suffer from data sparseness, we trained a naive Bayes model

for $P(R|U)$ (as, according to Bayes' rule, $P(U|R)$ is equal to $P(R|U)P(U)\frac{1}{P(R)}$, which, plugged in into formula (2), cancels out $\frac{1}{P(U)}$; further assuming the $P(R)$ is uniform, we can directly replace $P(U|R)$ with $P(R|U)$ here). On the language side (the variable $U$ in the model), we used n-grams over Japanese characters (we attempted tokenisation of the REs into words, but found that using characters worked just as well in the held-out set).

**P(I)** The prior $P(I)$ is the posterior of the previously computed increment. In the first increment, it can simply be set to a uniform distribution. Here, we apply a more informative prior based on saliency. We learn a *context model* which is queried when the first word begins, taking information about the context immediately before the beginning of the RE into account, producing a distribution over objects, which becomes $P(I)$ of the first increment in the RE. The context model itself is a simple application of the SIUM, where instead of being a word, $U$ is a token that represents saliency. The context model thus learns what properties are important to the pre-RE context and provides an up-to-date distribution over the objects as a RE begins.

### 5.1 Example

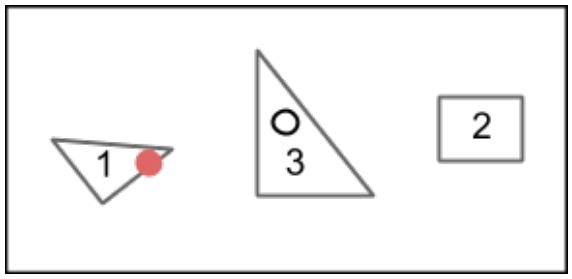


Figure 3: Example scene with two triangles and one square, 1 is being looked at by the SV, 3 was recently moved and the mouse pointer is still over it.

We will now give a simple example of how the model is applied to the REX data using a subset of the above properties for the RE *small triangle*. Table 3 shows a simple normalised co-occurrence count of how many times properties were observed as belonging to a referred object (the basis for $P(U|R)$). Figure 3 shows the current toy scene, and Table 4 shows the properties that each object in the scene has during the utterance. Table 5 shows the full application of the model by summing over the properties for the product $P(U|R)P(R|I)$ and multiplying by the prior $P(I)$, the posterior of the previous step. Included in this example is how the initial prior is computed from the context model.

| property | *small* | *triangle* | *square* | ⟨*context*⟩ |
|---|---|---|---|---|
| **small** | .87 | .02 | .4 | .04 |
| **big** | .01 | .08 | .02 | .05 |
| **triangle** | .04 | .88 | .01 | .09 |
| **square** | .04 | .01 | .9 | .09 |
| **left** | .06 | .07 | .06 | .09 |
| **center** | .04 | .03 | .04 | .07 |
| **right** | .04 | .06 | .05 | .03 |
| **most_gazed** | .07 | .09 | .07 | .6 |
| **recent_move** | .03 | .1 | .04 | .56 |
| **mouse_pointed** | .08 | .05 | .06 | .71 |

Table 3: Applications of $P(U|R)$, for some values of $U$ and $R$; we assume that this model is learned from data (rows are excerpted from a larger distribution over all the words in the vocabulary)

| property | 1 | 2 | 3 |
|---|---|---|---|
| **small** | 0.25 | 0.33 | 0 |
| **big** | 0 | 0 | 0.2 |
| **triangle** | 0.25 | 0 | 0.2 |
| **square** | 0 | 0.33 | 0 |
| **left** | 0.25 | 0 | 0 |
| **center** | 0 | 0 | 0.2 |
| **right** | 0 | 0.33 | 0 |
| **most_gazed** | 0.25 | 0 | 0 |
| **recent_move** | 0 | 0 | 0.2 |
| **mouse_pointed** | 0 | 0 | 0.2 |

Table 4: $P(R|I)$, for our example domain. The probability mass is distributed over the number of properties that a candidate object actually has.

Before the RE even begins, the prior saliency yields that 3 is the most likely object to be referred; it was the most salient in that it was the most recently moved object and the mouse pointer was still over it. However, initial prior information alone is not enough to resolve the intended object; for that the RE is needed. After the word *small* is uttered, 1 is the most likely referred object. After *triangle*, 1 remains the highest in the distribution. With the RE alone, in this case there would have been enough information to infer that 1 was the referred object, but adding the prior information provided additional evidence.

| $I$ | $U$ | $\Sigma\, P(U\|R) * P(R\|I)$ | $P(I\|U)$ |
|---|---|---|---|
| 1 | ⟨*context*⟩ | $.25(.04 + .09 + .09 + .6)$ | .37 |
| 2 | | $.33(.04 + .09 + .03)$ | .095 |
| 3 | | $.2(.05 + .09 + .07 + .56 + .71)$ | .535 |
| 1 | *small* | $.25(.87 + .04 + .06 + .07)$ | .65 |
| 2 | | $.33(.87 + .04 + .04)$ | .2 |
| 3 | | $.2(.01 + .04 + .04 + .03 + .08)$ | .15 |
| 1 | *triangle* | $.25(.02 + .88 + .07 + .09)$ | .81 |
| 2 | | $.33(.02 + .01 + .06)$ | .028 |
| 3 | | $.2(.08 + .88 + .03 + .1 + .05)$ | .162 |

Table 5: Application of RE *small triangle*, where 1 is the referred object

**Evaluation Metrics** We report results of our evaluation in referential accuracy on utterances that were annotated as referring to a single object (references to group objects is left for future work). Going beyond Iida et al. (2011), our model computes a resolution hypothesis incrementally; for the performance of this aspect of the system we followed previously used metrics for evaluation (Schlangen et al., 2009; Kennington et al., 2013):

**first correct**: how deep into the RE does the model predict the referent for the first time?

**first final**: how deep into the RE does the model predict the correct referent and keep that decision until the end?

**edit overhead**: how often did the model unnecessarily change its prediction (the only *necessary* prediction happens when it first makes a correct prediction)?

We compare non-incremental results to three evaluations performed in Iida et al. (2011), namely when Ling is used alone, Ling+TaskSP used together, and Ling+TaskSp+Gaze. Furthermore, they show results of models where a separate part handled REs that used pronouns, as well as a part that handled the non-pronoun REs, and a combined model that handled both types of expressions.

## 6 Results

### 6.1 Reference Resolution

Results of our evaluation are shown in Figure 4. The SIUM model performs better than the combined approach of Iida et al. (2011), and performs better than their separated model—when not including gaze (there is a significant difference between SIUM and the separated models for Ling+TaskSp, though

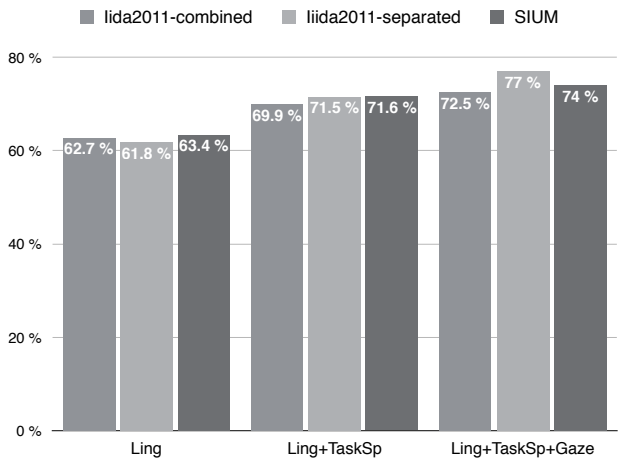


Figure 4: Comparison of model accuracies; our SIUM approach generally performs better over the combined model presented in Iida et al., (2011)

SIUM only got one more correct than the separated model). This is a welcome result, as it shows that our very simple incremental model that uses a basic classifier is comparable to a non-incremental approach that uses a more complicated classifier. It further shows that the SIUM model is robust to using TaskSp and Gaze features as properties, as long as those features are available immediately before the RE begins, or during the RE.

The best-performing approach is the Iida2011-separated model with gaze. This is the case for several reasons: first, their models use features that are not available to our incremental model (e.g., their model uses 14 gaze features, some of which were based on the entire RE, ours only uses 4 properties). Second, and more importantly, separated models means less feature confusion: in Iida et al. (2011) (Section 5.2), the authors give a comparison of the most informative features for each model; task and gaze features were prominent for the pronoun model, whereas gaze and language features were prominent for the non-pronoun model. We also tested SIUM under separated conditions to better compare with the approaches presented here. The separated models, however, did not improve. This, we assume, is because the model grounds *language* with properties (see Discussion below). An interactive dialogue system might not have the luxury of choosing between two models at runtime. We assume that a model that can sufficiently handle both

| | 1-5 | 6-8 | 9-14 |
|---|---|---|---|
| first correct (% into RE) | 35.47 | 22.34 | 14.8 |
| first final (% into RE) | 69.0 | 49.85 | 48.0 |
| edit overhead (all lengths) | 0.88% | | |
| never correct (all lengths) | 5.5% | | |

Table 6: Incremental results for SIUM, numbers represent % into RE.

types of utterances is to be preferred to one that doesn't.

### 6.2 Incremental Behaviour

Table 6 shows how our model fares using the incremental metrics described earlier. (As this has not been done in Iida et al. (2011), direct comparison is not possible.) For the evaluation, REs are binned into short, normal, and long (1-5, 6-8, 9-14 characters, respectively, based on what the average numbers of words in REs in this corpus is), to make relative statements ("% into the utterance") comparable.

Ideally, a system would make the first correct decision as early as possible without changing that decision. The results in the table show a respectable incremental model; on average it picks the right object early, with some edit overhead (making unnecessary changes in its prediction), finally fixing on a final decision before the end of the RE with low edit overhead, meaning it rarely changes its mind once it has made a decision. In some cases, SIUM never guessed the correct object, labeled *never correct* in the table. These incremental results are consistent with previous work for the SIUM; overall, the model is stable across the RE.

### 6.3 Discussion

Despite being very simple, there is an important difference that allows SIUM to improve over previous work. It learns to connect object properties selected for verbalisation to ways of verbalising them, and forms a stochastic expectation about which properties might be selected for verbalisation (namely, those that are present). This represents a type of *grounding* (Harnad, 1990; Roy, 2005).[4] In terms of the SIUM formalism, the link between object and words is mediated by the properties the object has

[4]Not to be confused with *building common ground* (Clark, 1996) which is also referred to as *grounding*.

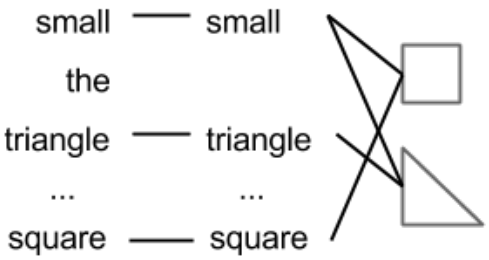


Figure 5: Words that describe objects are linked via a hand-coded `compatibility` function; links from words to multiple properties can exist, provided it is coded.

and by a stochastic process of associating words with properties. Figure 6 visualises this: each word has a stochastic connection between each property and objects have a set of properties. The property names are arbitrary as long as they are consistent. In contrast, previous work in RR (Iida et al., 2011; Chai et al., 2014) used a hand-coded concept-labeled semantic representation and checked if aspects of the RE match that of a particular object. If so, a binary *compatibility* feature was set. Figure 5 shows this; words can only link to objects via hand-crafted rules (e.g., the word or FOL predicate and property string must match). By the way SIUM uses properties, it can also perform (exophoric) pronoun resolution, deixis (the mouse pointer) and definite descriptions, in a single framework. This is a nice feature of the model: adding additional modalities does not require model reformulation.

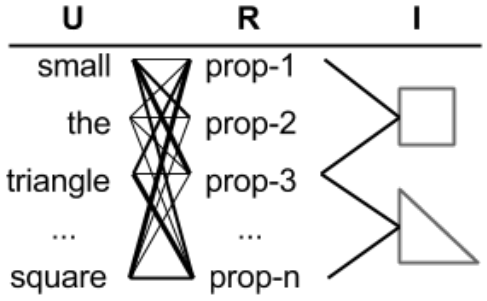


Figure 6: Words that describe objects are linked via properties stochastically: thicker lines between $U$ and $R$ represent higher probabilities. The lines between $R$ and $I$ denote a property belonging to an object. The cardinality of $U$ does not equal $R$.

Incorporating saliency information via a context model is also a nice feature of the model. In this paper, we computed the initial $P(I)$ using a context model instantiated by SIUM. By considering only this saliency information, the context model can predict the referred object in 41% of the cases. It also learned which properties were important for saliency (that is, these are the properties that the model would most likely select): `recently_fixated`, `most_gaze_at`, `longest_gazed_at`, `prev_ref`, as might be expected. In less than 2% of the cases, the context model referred to the correct object, but was wrongly "overruled" when processing the corresponding RE.

There were shortcomings, however. In previous work, it was shown that SIUM performed well when REs contained pronouns (see Kennington et al. (2013), experiment 2). However, in the current work we observed that REs with pronouns were more difficult for the model to resolve than the model presented in Iida et al. (2011). We surmise that SIUM had a difficult time grounding certain properties, as the Japanese pronoun *sore* can be used anaphorically or demonstratively in this kind of context (i.e., sometimes *sore* refers to previously-manipulated objects, or objects that are newly identified with a mouse pointer over them); the model presented in Iida et al. (2011) made more use of contextual information when pronouns were used, particularly in the combined model which incorporated gaze information, as shown above.

## 7 Conclusion

The SIUM is a model of RR that grounds language with the world, works incrementally, can incorporate modalities such as gaze and deixis, and can resolve multiple kinds of RRs in a single framework. This paper represents the natural next step in evaluating SIUM in a setting that was less constrained and more interactive, with added knowledge that it can work in more than one language.

There is more to be tested for SIUM. A common form of RR happens collaboratively over multiple utterances (Clark and Wilkes-Gibbs, 1986; Heeman and Hirst, 1995), SIUM has only been tested on isolated REs. Though SIUM required fewer features (realised as properties) than previous work, those properties still need to be computed. We leave for future work investigation of a version of the model that can ground language with raw(er) information from the world (e.g., vision information), eliminating the need to determine properties.

**Acknowledgements** Thanks to the anonymous reviewers for their excellent comments and suggestions.

## References

Gregory Aist, James Allen, Ellen Campana, Carlos Gomez Gallo, Scott Stoness, and Mary Swift. 2007. Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. In *Pragmatics*, volume 1, pages 149–154, Trento, Italy.

Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 33–40, Bielefeld, Germany.

Lin Chen and Barbara Di Eugenio. 2012. Co-reference via Pointing and Haptics in Multi-Modal Dialogues. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 523–527. Association for Computational Linguistics.

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

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

Kotaro Funakoshi, Mikio Nakano, Takenobu Tokunaga, and Ryu Iida. 2012. A Unified Probabilistic Approach to Referring Expressions. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 237–246, Seoul, South Korea, July. Association for Computational Linguistics.

Dilek Hakkani-tür, Malcolm Slaney, Asli Celikyilmaz, and Larry Heck. 2014. Eye Gaze for Spoken Language Understanding in Multi-Modal Conversational Interactions. In *ICMI*.

Stevan Harnad. 1990. The Symbol Grounding Problem. *Physica D*, 42:335–346.

Peter a. Heeman and Graeme Hirst. 1995. Collaborating on Referring Expressions. *Computational Linguistics*, 21(3):32.

Ryu Iida, Shumpei Kobayashi, and Takenobu Tokunaga. 2010. Incorporating Extra-linguistic Information into Reference Resolution in Collaborative Task Dialogue. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistic*, pages 1259—-1267, Uppsala, Sweden.

Ryu Iida, Masaaki Yasuhara, and Takenobu Tokunaga. 2011. Multi-modal Reference Resolution in Situated Dialogue by Integrating Linguistic and Extra-Linguistic Clues. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 84–92.

Ellen a. Isaacs and Herbert H. Clark. 1987. References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1):26–37.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, pages 133–142.

Andrew Kehler. 2000. Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction. In *AAAI 00, The 15th Annual Conference of the American Association for Artificial Intelligence*, pages 685–689.

Casey Kennington and David Schlangen. 2013. Situated incremental natural language understanding using Markov Logic Networks. *Computer Speech & Language*, pages –.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2013. Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *SIGdial 2013*.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2014. Situated Incremental Natural Language Understanding using a Multimodal, Linguistically-driven Update Model. In *CoLing 2014*.

Hatim Khouzaimi, Romain Laroche, and Fabrice Lefevre. 2014. An easy method to make dialogue systems incremental. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 98–107, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Spyros Kousidis, Casey Kennington, and David Schlangen. 2013. Investigating speaker gaze and pointing behaviour in human-computer interaction with the mint.tools collection. In *SIGdial 2013*.

Changsong Liu, Rui Fang, Lanbo She, and Joyce Chai. 2013. Modeling Collaborative Referring for Situated Referential Grounding. In *Proceedings of the SIGDIAL 2013 Conference*, pages 78–86, Metz, France, August. Association for Computational Linguistics.

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions. In *Aaai*. AAAI Press.

Teruhisa Misu, Antoine Raux, Ian Lane, and Moffett Field. 2014. Situated Language Understanding at 25 Miles per Hour. In *SIGdial 2014*, pages 22–31, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Andreas Peldszus, Okko Buß, Timo Baumann, and David Schlangen. 2012. Joint Satisfaction of Syntactic and Pragmatic Constraints Improves Incremental Spoken Language Understanding. In *Proceedings of the 13th EACL*, pages 514–523, Avignon, France, April. Association for Computational Linguistics.

Luis Pineda and Gabriela Garza. 2000. A Model for Multimodal Reference Resolution. *Computational Linguistics*, 26:139–193.

Massimo Poesio and Renata Vieira. 1997. A Corpus-Based Investigation of Definite Description Use. *Computational Linguistics*, 24(2):47, June.

Zahar Prasov and Joyce Y. Chai. 2008. What's in a gaze? In *Proceedings of the 13th international conference on Intelligent user interfaces - IUI '08*, page 20.

Zahar Prasov and Joyce Y Chai. 2010. Fusing Eye Gaze with Speech Recognition Hypotheses to Resolve Exophoric References in Situated Dialogue. In *Emnlp 2010*, number October, pages 471–481.

Deb Roy. 2005. Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciences*, 9(8):389–396, August.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on EACL 09*, 2(1):710–718.

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental Reference Resolution: The Task, Metrics for Evaluation, and a Bayesian Filtering Model that is Sensitive to Disfluencies. In *Proceedings of the 10th SIGdial*, number September, pages 30–37, London, UK. Association for Computational Linguistics.

David Schlangen. 2004. *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D. thesis, University of Edinburgh.

William Schuler, Stephen Wu, and Lane Schwartz. 2009. A Framework for Fast Incremental Interpretation during Speech Decoding. *Computational Linguistics*, 35(3):313–343.

Alexander Siebert and David Schlangen. 2008. A Simple Method for Resolution of Definite Reference in a Shared Visual Context. In *Proceedings of the 9th SIGdial*, number June, pages 84–87, Columbus, Ohio. Association for Computational Linguistics.

Gabriel Skantze and Anna Hjalmarsson. 1991. Towards Incremental Speech Production in Dialogue Systems. In *Word Journal Of The International Linguistic Association*, pages 1–8, Tokyo, Japan, September.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on EACL 09*, (April):745–753.

Philipp Spanger, Masaaki Yasuhara, Ryu Iida, Takenobu Tokunaga, Asuka Terai, and Naoko Kuriyama. 2012. REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources and Evaluation*, 46(3):461–491, December.

Michael J Spivey, Michael K Tanenhaus, Kathleen M Eberhard, and Julie C Sedivy. 2002. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45(4):447–481.

Michael K Tanenhaus and Michael J Spivey-Knowlton. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632.

Takenobu Tokunaga, Ryu Iida, Asuka Terai, and Naoko Kuriyama. 2012. The REX corpora : A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 422–429.

David Traum, David Devault, Jina Lee, Zhiyang Wang, and Stacy Marsella. 2012. Incremental dialogue understanding and feedback for multiparty, multimodal conversation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7502 LNAI, pages 275–288. Springer.