The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues

Tokunaga Takenobu[†], **Iida Ryu**[†], **Terai Asuka**[‡] **and Kuriyama Naoko**[§]

[†]Department of Computer Science [‡]Global Edge Institute [§]Department of Human System Science
Tokyo Institute of Technology
2-12-1 Ôokayama, Meguro-ku, Tokyo, 152-8552, Japan
take@cl.cs.titech.ac.jp

Abstract

This paper describes a collection of multimodal corpora of referring expressions, the REX corpora. The corpora have two notable features, namely (1) they include time-aligned extra-linguistic information such as participant actions and eye-gaze on top of linguistic information, (2) dialogues were collected with various configurations in terms of the puzzle type, hinting and language. After describing how the corpora were constructed and sketching out each, we present an analysis of various statistics for the corpora with respect to the various configurations mentioned above. The analysis showed that the corpora have different characteristics in the number of utterances and referring expressions in a dialogue, the task completion time and the attributes used in the referring expressions. In this respect, we succeeded in constructing a collection of corpora that included a variety of characteristics by changing the configurations for each set of dialogues, as originally planned. The corpora are now under preparation for publication, to be used for research on human reference behaviour.

Keywords: referring expression, multimodal, corpus, situated dialogue, puzzle solving

1 Introduction

A referring expression is a linguistic device referring to a certain object. Its proper use plays a key role in realising smooth communication, such as between humans and computers in a multimodal setting. Researchers have tackled problems related to referring expressions from two commonly held viewpoints: their understanding and their generation.

The understanding of referring expressions, i.e. referent identification, has been actively studied in anaphora resolution research, the goal of which is to identify antecedents of anaphors in a text by using contextual information (Mitkov, 2002). In contrast, referring expression generation has been studied for distinguishing the target object from distractors in a given situation, by generating a concise and unambiguous referring expression (Dale and Reiter, 1995). Although early work dealt with isolated expressions in a static environment, the research focus has shifted to generating referring expressions in a dynamic and multimodal environment, which more closely resembles to the real world. In such environments, extra-linguistic information as well as linguistic information both play important roles for smooth communication. Multimodal corpora of referring expressions are therefore indispensable in order to deepen our understanding of and further research on referring expressions in more realistic settings. This paper describes the details of our corpora named the REX corpora, consisting of referring expressions used in collaborative problem solving dialogues where two participants collaboratively solve geometric puzzles. All referring expressions are annotated in terms of its referent and various attributes. The corpora also include extra-linguistic information such as participants' actions and eye-gaze.

The paper is structured as follows. Related work is described in section 2, and then we sketch out the REX corpora in section 3. Section 4 describes construction of the corpora and section 5 provides their analysis. Finally section 6 concludes the paper.

2 Related work

Over the last decade, with a growing awareness that referring expressions are frequently used in the context of a collaborative task (Clark and Wilkes-Gibbs, 1986; Heeman and Hirst, 1995), a number of corpora have been constructed in order to study referring expressions in such a domain.

The COCONUT corpus (Di Eugenio et al., 2000) is collected from keyboard-dialogues where the participants collaborate on a simple 2-D design task, i.e. buying and arranging furniture for two rooms. The COCONUT corpus has rich annotations at the linguistic and the intentional level but does not include any extra-linguistic information such as physical actions by participants.

More recent work has mainly concentrated on seeking to overcome the shortcoming of domain complexity. The QUAKE corpus (Byron and Fosler-Lussier, 2006) as well as its successor, the SCARE corpus (Stoia et al., 2008) are based on interactions captured in a 3-D virtual world where two participants collaboratively carry out a treasure hunting task. Although the task environment is complex, the participant actions are very limited relative to the task environment.

The JCT (Joint Construction Task) corpus was created based on the interactions of two participants collaboratively constructing a puzzle (Foster et al., 2008). The setting of the experiment is quite realistic and natural in that both participants achieve the goal on an equal footing. They also recorded participants' eye-gaze during the interaction. While the authors noted that the "transcribed speech was precisely time-aligned with all the visual and action com-

corpus	puzzle	hint	language	eye-gaze
T2008-08	Tangram	yes	Japanese	no
T2009-11	Tangram	yes	Japanese	yes
N2009-11	Tangram	no	Japanese	~~no~~ yes
P2009-11	Polyomino	yes	Japanese	yes
D2009-11	Double-Tangram	yes	Japanese	yes
T2010-03	Tangram	yes	English	no

Table 1: Dialogue configurations of the REX corpora

ponents of the construction process”, they do not provide further details on recorded information.

In contrast to these previous corpora, our REX corpora records a wide range of information useful for studying human reference behaviour in a situated dialogue. While the domain of our corpora is simple compared to the QUAKE and SCARE corpora, we allowed a comparatively large flexibility in the actions necessary for achieving the task goal. In addition, human behaviour in geometric puzzle solving has been extensively studied in Cognitive Science (Baran et al., 2007; Evans et al., 2011). We can utilise the accumulated research results in this field.

Our task setting is similar to that of the JCT corpus except that we assign different roles to participants in order to efficiently elicit referring expressions. In addition, we collected dialogues under various configurations with respect to puzzle type, hinting and language.

3 Overview of the corpora

The REX corpora consist of six corpora as shown in Table 1. Each corpus is constructed based on collaborative problem solving dialogues with different configurations, defined in terms of the following factors.

- Puzzle (Tangram, Polyomino and Double-Tangram)
- Hinting (with or without hints)
- Language (Japanese and English)

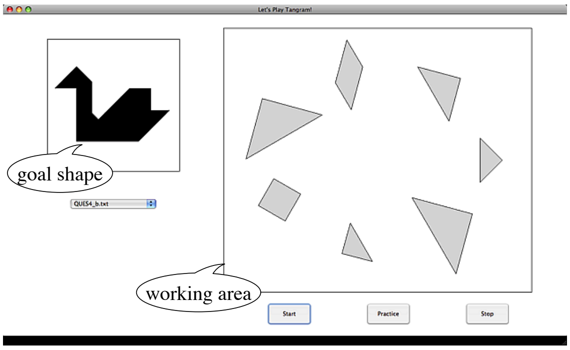


Figure 1: Tangram puzzle

We prepared three types of geometrical puzzles aimed to elicit different kinds of referring expressions depending on the puzzle type. The goal of the puzzles is to construct a given goal shape by arranging puzzle pieces (see Appendix for the goal shapes used in the data collection.). The puzzle pieces of Tangram consist of seven simple shapes: five triangles, a square and a parallelogram as shown in Figure 1.

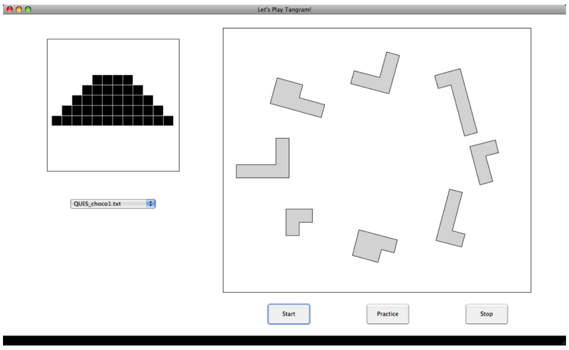


Figure 2: Polyomino puzzle

In contrast, the puzzle pieces of Polyomino are irregular shapes made of several unit squares (Figure 2). The pieces are more difficult to name in Polyomino than in Tangram, thus we expected more figurative expressions in dialogues for solving Polyomino.

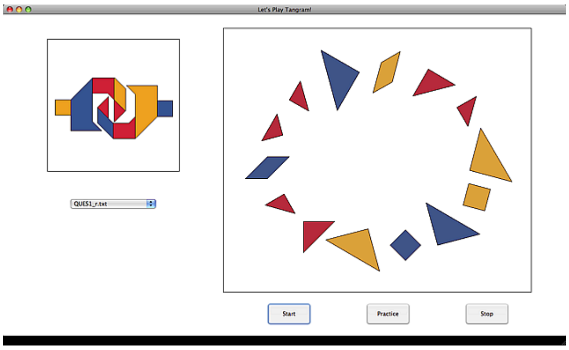


Figure 3: Double-Tangram puzzle

In the Double-Tangram puzzle, the participants are instructed to create a goal shape by using two sets of Tangram pieces with different colours (red, yellow and blue) as shown in Figure 3. This setting increases ambiguities since the number of pieces with the same attributes doubles. To remedy ambiguity, we introduced a colour attribute for the pieces.

In order to prevent the participants from getting into deep thought and keeping silent while solving a puzzle, we provided hints by showing a correct piece position in the goal shape except for the N2009-11 corpus (See Table 1).

The eye-gaze of both participants was captured in synchronisation with utterances except for the T2008-08 and T2010-03 corpora. This is because eye trackers were yet available at the time these dialogues were collected.

Our collected corpora are all Japanese except for the T2010-3 corpus which contains dialogues of English speakers.

Table 1 summarises the dialogue configurations of the REX corpora. The total number of annotated referring expressions is 8,859. Our previous research already utilised some of these corpora, e.g. T2008-08 in (Spanger et al., 2009a; Spanger et al., 2009b; Spanger et al., 2010b; Spanger et al., 2010a; Iida et al., 2010), T2008-8 and T2010-03 in (Tokunaga et al., 2010), T2009-11 in (Terai et al., 2011) and N2009-11 in (Kuriyama et al., 2011).

4 Corpus construction

4.1 Data collection

We recruited subjects as pairs of friends and colleagues. Each pair was instructed to solve a puzzle collaboratively. With the aim of recording the precise position of every piece and every action the participants made during the solving process, we implemented a puzzle simulator (Figure 1, 2, 3) in which the pieces can be moved, rotated and flipped with simple mouse operations on a computer display. The simulator displays two areas: a goal shape area and a working area where the pieces can be manipulated with their movements shown in real time.

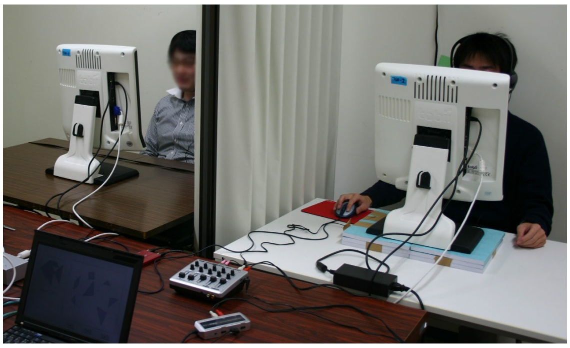


Figure 4: Picture of the experiment setting

We assigned a different role to each participant of a pair: one acts as a *solver* and the other as an *operator*. The operator has a mouse for manipulating puzzle pieces, but does not have a goal shape on the screen. The solver has a goal shape on the screen but does not have a mouse. This setting naturally leads to a situation where given a certain goal shape, the solver thinks of the necessary arrangement of the pieces and gives instructions to the operator where to move them, while the operator manipulates the pieces with the mouse according to the solver's instructions.

Each pair of participants sat side by side as shown in Figure 4 and solved 4 trials for the Tangram and Polyomino puzzles and 6 trials for the Double-Tangram puzzle. Each participant had his/her own computer display showing the shared working area in real time. A room-divider screen was set between the solver and operator to prevent the operator from seeing the goal shape on the solver's screen, and to restrict their interaction to speech and the shared working area on the computer display. We did not constrain the contents of their dialogues. The participants exchanged roles after half of the assigned trials. The order of the puzzle trials is the same for all pairs.

Before starting the first trial as the operator, each participant had a short training exercise in order to learn how to manipulate pieces with the mouse. The initial arrangement of the pieces was randomised each time. We set a time limit of 15 minutes for the completion of each trial for the Tangram, and 10 minutes for the Polyomino and Double-Tangram. These limits were decided based on preliminary experiments.

In order to prevent the solver from getting into deep thought and keeping silent, the Tangram simulator is designed to give a hint every five minutes by showing a correct piece position in the goal shape area. After 10 minutes have passed, a second hint is provided, while the previous hint disappears. Since Polyomino is more difficult than Tangram, the Polyomino simulator provides a correct piece position from the beginning and accumulates further correct piece positions every two and half minutes. The Double-Tangram simulator shows partially solved goal shapes from the beginning and adds no further hint during the dialogues. A trial ends when the goal shape is completed or the time is up. Utterances by the participants are recorded separately in stereo through headset microphones in synchronisation with the position of the pieces and the mouse operations and eye gaze of both participants. Piece positions and mouse actions were automatically recorded by the simulator at intervals of 1/65 second. Eye gaze was captured by the Tobii T60 eye tracker at intervals of 1/60 second. The display size was $1,280 \times 1,024$ pixels and the distance between the display and each participant's eyes was maintained at about 45cm. We conducted the 9-point calibration for both participants before the trials.

Recent eye-tracking devices like Tobii have made it drastically easier to capture a subject's gaze positions. However, there still remain eye-tracking errors in the experiments. Following (Bard et al., 2007), we discarded the dialogues in which an erroneous duration of eye-tracking exceeds 30%[1] of the overall dialogue length.

dpr	: demonstrative pronoun, e.g. "the same one", "this", "that", "it"
dad	: demonstrative adjective, e.g. "that triangle"
dmn	: dummy noun, e.g. "*ue no* (the upper one)"
siz	: size, e.g. "the large triangle"
col	: colour, e.g. "the blue square"
typ	: type, e.g. "the square"
dir	: direction of a piece, e.g. "the triangle facing the left".
prj	: projective spatial relation, e.g. "the triangle to the left of the square"
tpl	: topological spatial relation, e.g. "the triangle near the square"
ovl	: overlap, e.g. "the small triangle under the large one"
act	: action on pieces, e.g "the triangle that you are holding now"
cmp	: complement, e.g. "the other one"
sim	: similarity, e.g. "the same one"
num	: number, e.g. "the two triangle"
rpr	: repair, e.g. "the big, no, small triangle"
err	: obvious erroneous expression, e.g. "the square" referring to a triangle
nest	: nested expression, e.g. "(the triangle to the left of (the square))"
meta	: metaphorical expression, e.g. "the leg", "the head"
nul	: no applicable attribute

Table 2: Attributes of referring expressions

4.2 Annotation

The recorded speech data was transcribed and the referring expressions were annotated with the multimodal annotation

[1] There are four exceptional dialogues in which the erroneous duration exceeds 30% in T2009-11 and N2009-11. But they are still less than 40%.

step	(A_1, A_2)	(A_1, A_3)
(1) expression identification (β)	0.67	0.75
(2) referent identification (κ)	0.81	0.93
(3) attributes assignment (κ)	0.86	0.87

Table 3: Agreement analysis of 9 dialogues in T2009-11

tool ELAN[2]. We limited annotations to expressions referring to a puzzle piece or a set of puzzle pieces. The annotation of referring expressions is three-fold: (1) identification of the span of expressions, (2) identification of their referents, and (3) assignment of a set of attributes to each referring expression. The annotation guidelines basically follow the ones described in (Spanger et al., 2010b). Referents of an expression are represented as a sequence of piece IDs. An expression without a definite referent, i.e. referring to a class instead of an instance, is marked with a prefix, followed by a sequence of possible piece IDs. Attributes of expressions are shown in Table 2. Note that multiple attributes can be assigned to an expression. The nul attribute is assigned to an expression if it has no applicable attributes. We analysed inter-annotator agreement on each of the above steps for 9 out of 27 dialogues from the T2009-11 corpus[3]. We employed three independent annotators, A_1, A_2 and A_3 to make two pairs (A_1, A_2) and (A_1, A_3)[4]. Each pair worked on 4 and 5 dialogues respectively. To evaluate inter-annotator agreement of referent identification and attribute assignment (steps (2) and (3)), we adopted the κ-coefficient, which is considered as a *de facto* standard measure in evaluating categorical matching of corpus annotation (Carletta, 1996). In contrast, for considering the degree of overlap between text spans by different annotators, we adopted the β-coefficient (Artstein and Poesio, 2005) for the expression identification (step (1)). While the κ-coefficient makes a binary decision for matching, the β-coefficient uses a continuous value to represent a matching degree. We follow (Foster and Oberlander, 2007) to set up the matching values as follows[5].

1	: exact match
2/3	: one span subsumes other span
1/3	: two spans overlap
0	: mismatch

Table 3 shows the agreement coefficient of (1) expression identification, (2) referent identification and (3) attribute assignment. The results indicate that we obtained fairly stable annotations. The rest of the dialogues were annotated by a single annotator and validated by one of the authors to finalise the corpora.

tier	meaning
OP-UT	utterances (operator)
SV-UT	utterances (solver)
OP-REX	referring expressions (operator)
OP-Ref	referents of OP-REX
OP-Attr	attributes of OP-REX
SV-REX	referring expressions (solver)
SV-Ref	referents of SV-REX
SV-Attr	attributes of SV-REX
Action	action on a piece
Target	the target piece of Action
Mouse	the piece on which the mouse is hovering
OP-GZE-P	fixation point (operator)
OP-GZE-N	fixation piece (operator)
SV-GZE-P	fixation point (solver)
SV-GZE-N	fixation piece (solver)

∗ Indentation of tier denotes parent-child relations.

Table 4: The ELAN tiers

corpus	#pairs	#dialg.	#valid	#succ.	comp. time (SD)
T2008-08	6	24	24	21	10:42 (3:16)
T2009-11	8	32	27	23	9:43 (3:32)
N2009-11	5	20	8	4	13:28 (2:48)
P2009-11	7	28	24	24	6:07 (1:33)
D2009-11	7	42	24	23	5:53 (2:08)
T2010-03	6	24	24	10	12:47 (3:34)
T2008-08+ T2009-11	13	56	51	46	10:11 (3:25)

Table 5: Number of dialogues and average completion time

These annotations were then merged with extra-linguistic information, i.e. action data from the puzzle simulator and eye-gaze data from the eye tracker, using the ELAN annotation tool. The available action information from the simulator consists of the action on a piece, the coordinates of the mouse cursor and the position of each piece in the working area. Actions are annotated as a time span labeled with an action name ("move", "rotate" or "flip") and its target piece ID, and mouse cursor positions are annotated as a time span labeled with a piece ID for the piece under the mouse cursor during that span. The position of pieces is updated and recorded with a time stamp when the position of any piece changes. Information about piece positions is not merged into the ELAN files and is kept in separate files.

Eye-gaze data is transformed into gaze fixation points and attended objects through the following calculation and added to the ELAN files. First we define gaze fixation using the Dispersion-Threshold Identification (I-DT) algorithm (Salvucci and Goldberg, 2000). Based on a rational that the eye movement velocity slows near fixations, the I-DT algorithm identifies fixations as clusters of consecutive gaze points within a particular dispersion. It has two parameters, the dispersion threshold which defines the maximum distance between gaze points belonging to the same cluster, and the duration threshold which constrains the minimum fixation duration. Considering experimental configurations, i.e. (1) the display size and its resolution, (2) the distance between the display and the subject's eyes, and (3) the eye-tracker resolution, we set the dispersion threshold to 16 pixels. Following (Richardson et al., 2007), we set the duration threshold to 100 msec. For a gaze fixation span, the centroid of fixations and the nearest piece ID to the centroid were added in separate ELAN tiers. As a result, we have 15 time-aligned ELAN tiers as shown in Table 4.

[2] http://www.lat-mpi.eu/tools/elan/

[3] We have already reported the agreement analysis of T2008-08 in (Spanger et al., 2010b).

[4] A_1 is one of the authors of this paper.

[5] (Foster and Oberlander, 2007) actually defined a distance that is dual to our matching value.

corpus	#utterances		#referring exp.	
	OP	SV	OP	SV
T2008-08				
Total	1,892	2,571	200	1,214
Ave.	78.8	107.1	8.3	50.6
SD	51.0	40.6	10.4	19.3
T2009-11				
Total	2,382	4,613	271	1,192
Ave.	88.2	170.9	10.0	44.1
SD	69.8	86.8	11.5	24.8
N2009-11				
Total	1,119	1,716	168	497
Ave.	139.9	214.5	21.0	62.1
SD	47.4	62.6	6.9	22.8
P2009-11				
Total	1,903	2,920	325	1,056
Ave.	79.3	121.7	13.5	44.0
SD	38.0	30.7	10.1	17.0
D2009-11				
Total	926	3,024	115	1,115
Ave.	38.6	126.0	4.8	46.5
SD	30.1	46.0	4.5	14.6
T2010-03				
Total	2,049	4,848	310	2,396
Ave.	85.4	202.0	12.9	99.8
SD	64.0	70.1	10.2	42.5
T2008-08+T2009-11				
Total	4,274	7,184	471	2,406
Ave.	83.8	140.9	9.2	47.2
SD	61.3	75.6	10.9	22.4

Table 6: Number of utterances and referring expressions

5 Details of the corpora

This section presents an analysis of the corpora, comparing various statistics across the corpora, which were collected in different configurations as described in section 3. Table 5, 6, 7 show the various statistics of each corpus. We add an extra row, labeled "T2008-08+T2009-11", which was made by merging corpora T2008-08 and T2009-11, because the configurations of these corpora are the same except for eye-gaze information, which is not mentioned in this article.

Table 5 shows the number of participant pairs (#pairs), the total number of collected dialogues (#dialg.), the number of valid dialogues which fulfil the condition on eye-tracking error rate (#valid), the number of successful trials among the valid dialogues (#succ.), and the average completion time of the trials with its standard deviation in parentheses (comp. time). The following analysis targets the data of valid dialogues.

Table 6 shows the number of utterances (#utterances) and referring expressions (#referring exp.) by operator (OP) and solver (SV) respectively. Since the number of valid dialogues is different across the corpora, we focus on the average number per dialogue.

Table 7 shows the occurrence of referring expression attributes. Again, we focus on the average number of attributes. In the succeeding subsections, we analyse the tendency of these statistics based on the difference of configurations during dialogue collection, i.e. puzzle type, hinting and language.

5.1 Puzzle type:

(T2008-08+T2009-11 vs. P2009-11 vs. D2009-11)

In order to see the difference among puzzle types, we compared the statistics of T2008-08+T2009-11 (Tangram), P2009-11 (Polyomino) and D2009-11 (Double Tangram). Notable differences can be observed in the average number of utterances and referring expressions in Table 6, and average number of attributes dpr, siz, typ, prj, act, rpr, err, nest and meta in Table 7. We conducted a one-way ANOVA with each of these values as the dependent variable and the puzzle type as the independent variable. For all values except for act and nest, the main effect was significant (operator utterances: $F(2,96) = 6.99$, $p < 0.01$, operator referring expressions: $F(2.96) = 5.04$, $p < 0.01$, dpr: $F(2.96) = 8.86$, $p < 0.01$, siz: $F(2.96) = 37.0$, $p < 0.01$, typ: $F(2.96) = 75.1$, $p < 0.01$, prj: $F(2.96) = 33.5, p < 0.01$, rpr: $F(2.96) = 11.6$, $p < 0.01$, err: $F(2.96) = 6.22$, $p < 0.01$, meta: $F(2.96) = 11.3$, $p < 0.01$). We assume these values are independent. But if we assume dependency among these and adopt a multiple comparison, the main effects of operator referring expressions and err disappear after applying the Bonferroni correction. In what follows, we assume independence among the values.

The average number of operator utterances in D2009-11 is significantly smaller than that of T2008-08+T2009-11 and P2009-11 at 1% significance level. Since the goal arrangement is almost shown as a hint from the beginning of a trial in the Double Tangram puzzle, the operators tend to just hear and perform the solver's instructions. This setting would suppress operator utterances. In contrast, in the other two puzzles, even though hints are provided, they are given in a step by step manner during the task; they place more weight on puzzle solving than manipulating puzzle pieces. This might have elicited more utterances from the operator. According to the number of utterances, the number of operator referring expressions in D2009-11 is smaller than in others, but only the difference between T2008-08+T2009-11 and D2009-11 was significant at 1% significance level.

We observe more frequent use of dpr, prj and meta in P2009-11 than in others. All differences are significant at 1% significance level. This can be explained as follows. The puzzle pieces of Polyomino have irregular shapes consisting of unit squares as shown in Figure 2, thus, the dialogue participants tend to use more pronouns (dpr), spatial relations (prj) and metaphorical expressions (meta) to refer to a piece.

Conversely, siz and typ are less used in P2009-11 than in others. The differences are significant at 1% significance level. This can be also explained by irregularity of the Polyomino piece shape. It is difficult to refer to the Polyomino pieces by usual figure types like "triangle" and "square". There are 4 cases using the typ attribute to refer to a piece in P2009-11. They includes two abstract expressions ("block" and "figure"), and two approximations ("rectangle" and "trapezium"). Mentioning a size of such irregular shapes would be also difficult.

The attribute typ is less used in D2009-11 than in T2008-08+T2009-11. The difference is significant at 1% significance level. This would be because the colour attribute (col) is available for D2009-11. Actually quite a lot of uses of col is observed in D2009-11, which compensates for uses of the typ attribute.

D2009-11 shows more uses of rpr and err. The difference is significant at 1% significance level against T2008-08+T2009-11, and at 5% significance level against P2009-11. This would be explained by that increased number of puzzle pieces induced more repairs (rpr) and errors (err).

5.2 Hinting: (T2008-08+T2009-11 vs. N2009-11)

Hinting effect was evaluated with the Tangram puzzle data by comparing T2008-08+T2009-11 (with hints) and N2009-11 (without hints). There seem be differences in the average completion time and the task success rate in Table 5, the average number of utterances and referring expressions in Table 6, and the average number of attributes dpr, dmn and siz in Table 7.

We conducted a χ^2-test for the success rate and a t-test for the other statistics. Among these statistics, we did not find significant differences for the number of solver referring expressions, the number of attributes dmn and siz. The difference of other values were significant (the completion time: $t(57) = 2.58,\ p < 0.05$, the success rate: $\chi^2(1) = 3.85,\ p < 0.05$, operator utterances: $t(57) = 2.47,\ p < 0.05$, solver utterances: $t(57) = 2.61,\ p < 0.05$, operator referring expressions: $t(57) = 2.95,\ p < 0.01$, dpr: $t(57) = 2.10,\ p < 0.05$). After applying the Bonferroni correction, however, significance of these differences disappear. We assume independence among the values hereafter.

It is natural to see a lower success rate for dialogues without any hints given, and thus also a longer time until puzzle completion. Note that since we set a time limit to 15 minutes for solving the Tangram puzzle, we considered the completion time as 15 minutes even when a puzzle was not actually solved within the time limit.

The increased number of operator utterances and referring expressions in N2009-11 suggests more active contribution by the operator to solving the puzzle, since the task is more difficult to solve without hints.

Compared to the results in the previous subsection, it is notable that the attribute distribution is more affected by the puzzle types than hinting.

5.3 Language: (T2008-08+T2009-11 vs. T2010-03)

The effect of language was evaluated by comparing T2008-08+T2009-11 (Japanese) and T2010-03 (English). The differences are observed in the task completion time and the task success rate in Table 5, the average number of operator and solver referring expressions in Table 6, and the average number of attributes dpr, dad, tpl, cmp, num and meta. According to the result of the χ^2-test for the success rate, the t-test for the completion time, and Welch's t-test for the others, the differences are significant except for operator referring expressions and meta (the completion time: $t(73) = -3.04,\ p < 0.01$, the success rate: $\chi^2(1) = 14.0,\ p < 0.01$, solver referring expressions: $t(29.1) = -5.71,\ p < 0.01$, dpr: $t(29.1) = -5.78,\ p < 0.01$, dad: $5(32.8) = -4.91,\ p < 0.01$, tpl: $t(26.4) = -4.10,\ p < 0.01$, cmp: $t(24.5) = -4.64,\ p < 0.01$, num: $t(25.6) = -2.93,\ p < 0.01$). After applying the Bonferroni correction, however, significance of the differences of the completion time and num disappear. Again, we assume independence among the values hereafter.

We have already reported on the comparison of the task completion time and the success rate between T2008-08 and T2010-03 elsewhere (Tokunaga et al., 2010). We added the T2009-11 Japanese Tangram data on top of the T2008-08 data in this analysis. The tendency is the same, meaning that the English subjects needed more time to solve the puzzles and their success rate is lower than that of Japanese. As noted in our previous paper, this would be attributed to the diversity of the English subjects in terms of their occupation and age. The familiarity with the Tangram puzzle would also affect the result. Many of the English subjects had no experience with this kind of geometric puzzle.

The English subjects tend to use more demonstratives (pronouns (dpr) and adjectives (dad)) than Japanese. This might be related to that Japanese tends to use ellipses rather than pronouns for referring to salient entities. We can not have decisive conclusion because the current corpora does not have annotated ellipses. Annotating ellipses in Japanese corpora and further analysis based on the annotation remain for future work.

More use of tpl, cmp and num by the English speakers is another notable difference. We could not find a particular explanation for the difference of tpl and num, but found that the English expressions assigned cmp were very skewed, i.e. 134 of 144 cases included "other". In contrast, Japanese expressions with cmp were more diverse and tended to consist of more than one word, e.g. "*mou hitotu no* (another)" and "*amatte iru* (remaining)". In short, English has a concise and convenient word "other" to denote complementary entities. This would be an explanation of the difference in usage of the cmp attribute between English and Japanese.

It is interesting to see that English speakers used the colour attribute (col) even though colour has no discrimination ability among pieces in our Tangram setting. Note that only the pieces of the Double Tangram puzzle are coloured. However, 30 out of 35 expressions with col were used by the same person. This would be attributed to personal characteristics.

6 Concluding remarks

We introduced a collection of multimodal corpora of referring expressions, the REX corpora, which were collected through situated dialogues for collaborative problem solving. The corpora have two notable features, namely (1) they include time-aligned extra-linguistic information such as participant actions and eye-gaze on top of linguistic information, (2) dialogues were collected with various configurations in terms of the puzzle type, hinting and language. We described our process for constructing the corpora, and the composition of the collection of corpora. We also presented an analysis of various statistics for the corpora with respect to the various configurations mentioned above. The analysis showed that the corpora have different characteristics in the number of utterances and referring expressions in a dialogue, the task completion time and the attributes used in the referring expressions. In this respect, we suc-

corpus	dpr	dad	dmn	siz	col	typ	dir	prj	tpl	ovl	act	cmp	sim	num	rpr	err	nest	meta	nul
T2008-08																			
Total	668	176	39	285	0	647	7	141	10	2	94	29	7	35	2	2	30	6	10
Ave.	27.8	7.3	1.6	11.9	0	27.0	0.3	5.9	0.4	0.1	3.9	1.2	0.3	1.5	0.1	0.1	1.3	0.3	0.4
SD	19.5	5.3	2.1	6.3	0	11.9	0.9	4.7	0.8	0.3	3.5	1.8	0.9	1.9	0.3	0.3	1.6	0.4	1.0
T2009-11																			
Total	674	152	60	358	0	691	2	92	4	0	48	25	2	16	4	2	7	12	0
Ave.	25.0	5.6	2.2	13.3	0	25.6	0.1	3.4	0.1	0	1.8	0.9	0.1	0.6	0.1	0.1	0.3	0.4	0
SD	26.5	6.5	3.3	6.6	0	10.9	0.3	2.9	0.5	0	1.8	1.1	0.3	0.7	0.4	0.3	0.6	1.1	0
N2009-11																			
Total	345	57	31	170	0	244	1	19	2	0	25	12	0	15	0	1	1	7	1
Ave.	43.1	7.1	3.9	21.3	0	30.5	0.1	2.4	0.3	0	3.1	1.5	0	1.9	0	0.1	0.1	0.9	0.1
SD	22.1	5.3	5.0	11.7	0	14.4	0.4	2.5	0.7	0	3.9	2.7	0	2.4	0	0.4	0.4	1.5	0.4
P2009-11																			
Total	1,087	104	52	52	0	4	0	287	0	7	99	14	0	30	9	4	14	59	0
Ave.	45.3	4.3	2.2	2.2	0	0.2	0	12.0	0	0.3	4.1	0.6	0	1.3	0.4	0.2	0.6	2.5	0
SD	21.3	3.1	3.0	3.4	0	0.4	0	7.5	0	0.7	2.8	1.1	0	1.7	0.6	0.6	1.1	4.0	0
D2009-11																			
Total	592	101	73	182	513	431	1	52	2	1	70	29	7	30	19	14	5	8	3
Ave.	24.7	4.2	3.0	7.6	21.4	18.0	0	2.2	0.1	0	2.9	1.2	0.3	1.3	0.8	0.6	0.2	0.3	0.1
SD	17.6	3.9	3.4	5.1	7.6	8.5	0.2	2.0	0.3	0.2	2.8	1.1	0.6	2.2	0.8	1.0	0.5	1.3	0.4
T2010-03																			
Total	1,835	374	0	422	35	725	2	132	40	7	48	144	0	79	7	5	24	22	10
Ave.	76.5	15.6	0	17.6	1.5	30.2	0.1	5.5	1.7	0.3	2.0	6.0	0	3.3	0.3	0.2	1.0	0.9	0.4
SD	42.9	10.5	0	10.2	4.5	13.7	0.3	4.2	2.1	0.7	1.8	5.6	0	4.2	0.6	0.4	1.2	1.8	0.9
T2008-08+T2009-11																			
Total	1,342	328	99	643	0	1,338	9	233	14	2	142	54	9	51	6	4	37	18	10
Ave.	26.3	6.4	1.9	12.6	0	26.2	0.2	4.6	0.3	0	2.8	1.1	0.2	1.0	0.1	0.1	0.7	0.4	0.2
SD	23.6	6.1	2.7	6.5	0	11.3	0.6	4.0	0.6	0.2	2.9	1.5	0.6	1.5	0.3	0.3	1.3	0.9	0.7

Table 7: Number of referring expression attributes

ceeded in constructing a collection of corpora that included a variety of characteristics by changing the configurations for each set of dialogues, as originally planned. The corpora are now under preparation for publication, to be used for research on human reference behaviour.

7 References

Ron Artstein and Massimo Poesio. 2005. Kappa^3 = alpha (or beta). Technical Report CSM-437, University of Essex.

Bahar Baran, Berrin Dogusoy, and Kursat Cagiltay. 2007. How do adults solve digital tangram problems? Analyzing cognitive strategies through eye tracking approach. In *HCI International 2007 - 12th International Conference - Part III*, pages 555–563.

Ellen Gurman Bard, Anne H. Anderson, Yiya Chen, Hannele B. M. Nicholson, Catriona Havard, and Sara Dalzel-Job. 2007. Let's you do that: Sharing the cognitive burdens of dialogue. *Journal of Memory and Language*, 57(4):616–641.

Donna K. Byron and Eric Fosler-Lussier. 2006. The OSU Quake 2004 corpus of two-party situated problem-solving dialogs. In *Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2006)*.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

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

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

Barbara Di Eugenio, Pamela W. Jordan, Richmond H. Thomason, and Johanna. D. Moore. 2000. The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human-Computer Studies*, 53(6):1017–1076.

Michael A. Evans, Eliot Feenstra, Emily Ryon, and David McNeill. 2011. A multimodal approach to coding discourse: Collaboration, distributed cognition, and geometric reasoning. *Computer-Supported Collaborative Learning*, 6(2):253–278.

Mary Ellen Foster and Jon Oberlander. 2007. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, 41(3–4):305–323.

Mary Ellen Foster, Ellen Gurman Bard, Markus Guhe, Robin L. Hill, Jon Oberlander, and Alois Knoll. 2008. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of 3rd Human-Robot Interaction*, pages 295–302.

Peter A. Heeman and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational Linguistics*, 21:351–382.

Ryu Iida, Shumpei Kobayashi, and Takenobu Tokunaga. 2010. Incorporating extra-linguistic information into reference resolution in collaborative task dialogue. In *Proceedings of 48th Annual Meeting of the Association for Computational Linguistics*, pages 1259–1267.

Naoko Kuriyama, Asuka Terai, Masaaki Yasuhara, Takenobu Tokunaga, Kimihiko Yamagishi, and Takashi Kusumi. 2011. Gaze matching of referring expressions in collaborative problem solving. In *Proceedings of Workshop on Dual Eye Tracking in CSCW (DUET 2011)*.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman.

Daniel C. Richardson, Rick Dale, and Michael J. Spivey. 2007. Eye movements in language and cognition: A brief introduction. In Monica Gonzalez-Marquez, Irene Mittelberg, Seana Coulson, and Michael J. Spivey, editors, *Methods in Cognitive Linguistics*, pages 323–344. John Benjamins.

Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications (ETRA '00)*, pages 71–78.

Philipp Spanger, Masaaki Yasuhara, Ryu Iida, and Takenobu Tokunaga. 2009a. A Japanese corpus of referring expressions used in a situated collaboration task. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 110 – 113.

Philipp Spanger, Masaaki Yasuhara, Ryu Iida, and Takenobu Tokunaga. 2009b. Using extra linguistic information for generating demonstrative pronouns in a situated collaboration task. In *Proceedings of PreCogSci 2009: Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference*.

Philipp Spanger, Ryu Iida, Takenobu Tokunaga, Asuka Teri, and Naoko Kuriyama. 2010a. Towards an extrinsic evaluation of referring expressions in situated dialogs. In John Kelleher, Brian Mac Namee, and Ielka van der Sluis, editors, *Proceedings of the Sixth International Natural Language Generation Conference (INGL 2010)*, pages 135–144.

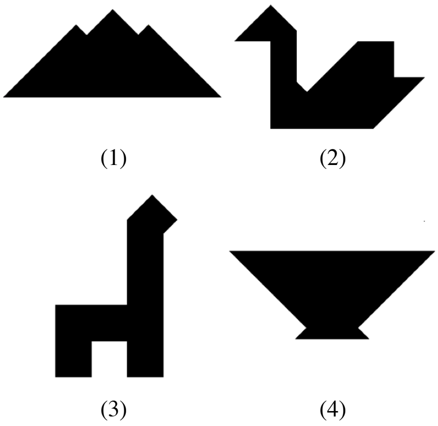
Philipp Spanger, Masaaki Yasuhara, Ryu Iida, Takenobu Tokunaga, Asuka Terai, and Naoko Kuriyama. 2010b. REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources and Evaluation*.

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. SCARE: A situated corpus with annotated referring expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 28–30.

Asuka Terai, Naoko Kuriyama, Masaaki Yasuhara, Takenobu Tokunaga, Kimihiko Yamagishi, and Takashi Kusumi. 2011. Using metaphors in collaborative problem solving: An eye-movement analysis. In *Proceedings of Workshop on Dual Eye Tracking in CSCW (DUET 2011)*.

Takenobu Tokunaga, Ryu Iida, Masaaki Yasuhara, Asuka Terai, David Morris, and Anja Belz. 2010. Construction of bilingual multimodal corpora of referring expressions in collaborative problem solving. In *Proceedings of 8th Workshop on Asian Language Resources*, pages 38–46.

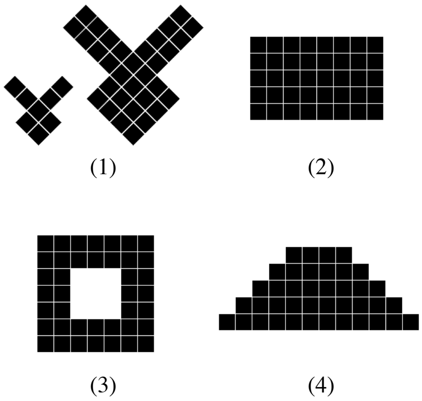
Appendix: Puzzle goal shapes



(1) (2)

(3) (4)

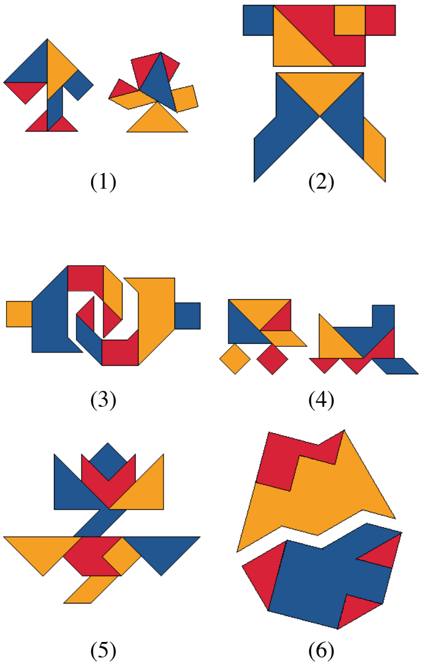
Tangram



(1) (2)

(3) (4)

Polyomino



(1) (2)

(3) (4)

(5) (6)

Double Tangram