Language Resources & Evaluation manuscript No.
(will be inserted by the editor)

REX-J: Japanese referring expression corpus of situated dialogs

Philipp Spanger[†] · **Masaaki Yasuhara**[†] ·
Ryu Iida[†] · **Takenobu Tokunaga**[†] ·
Asuka Terai[‡] · **Naoko Kuriyama**[§]

Received: ... / Accepted: date

Abstract Identifying objects in conversation is a fundamental human capability necessary to achieve efficient collaboration on any real world task. Hence the deepening of our understanding of human referential behaviour is indispensable for the creation of systems that collaborate with humans in a meaningful way. We present the construction of REX-J, a multi-modal Japanese corpus of referring expressions in situated dialogs, based on the collaborative task of solving the Tangram puzzle. This corpus contains 24 dialogs with over 4 hours of recordings and over 1400 referring expressions. We outline the characteristics of the collected data and point out the important differences from previous corpora. The corpus records extra-linguistic information during the interaction (e.g. the position of pieces, the actions on the pieces) in synchronization with the participants' utterances. This in turn allows us to discuss the importance of creating a unified model of linguistic and extra-linguistic information from a new perspective. Demonstrating the potential uses of this corpus, we present the analysis of a specific type of referring expression ("action-mentioning expression") as well as the results of research into the generation of demonstrative pronouns. Furthermore, we discuss some perspectives on potential uses of this corpus as well as our planned future work, underlining how it is a valuable addition to the existing databases in the community for the study and modeling of referring expressions in situated dialog.

Keywords multi-modal corpus · referring expressions · collaborative task · Japanese

1 Introduction

Referring expressions are a linguistic device referring to a certain object and play a fundamental role in smooth collaboration between humans and agents where physical operations are involved. Thus, research into the understanding and generation of referring expressions has been a critical research theme in order to develop more efficient and natural human-agent

† Department of Computer Science
‡ Global Edge Institute
§ Department of Human System Science
Tokyo Institute of Technology
E-mail: {philipp, yasuhara, ryu-i, take}@cl.cs.titech.ac.jp,
asuka@nm.hum.titech.ac.jp, kuriyama@hum.titech.ac.jp

DOI: 10.1007/s10579-010-9134-8

interaction. This has given rise to a substantial international research community within Computational Linguistics. The work presented in this paper is a contribution to the creation and utilization of resources for this field.

From the beginning, research in the field of analysis of referring expressions, by the nature of the problem it deals with (e.g. anaphora resolution), has been faced with the necessity of treating the broader textual context of referring expressions (Hobbs (1978); Grosz et al. (1995)). In contrast, in the field of language generation the expressions dealt with were mainly isolated expressions; hence early work in this area (Dale (1989); Dale and Reiter (1995)) focused largely on studying the generation of isolated referring expressions in a static environment of invariant situations (e.g. an image) and on how those expressions were able to distinguish the target object from the distractors.

Their simplifying assumption was that other factors outside the current (static) situation, such as the context of interaction, would not have an impact on the formulation and understanding of a referring expression. Based on these same assumptions and in order to allow a unified evaluation of such algorithms, the TUNA corpus was recently developed at Aberdeen University (van Deemter (2007)). It is the most extensive collection of referring expressions to date, containing roughly 2,000 expressions. At the same time, it has the limitation of only taking into account individual expressions in an interaction-free setting. Another recent corpus that has a similar limitation is the GRE3D3 corpus (Dale and Viethen (2009)). This corpus is focused on recording how humans use relational referring expressions and is significantly smaller (about a third of the size of the TUNA corpus). Both the TUNA and GRE3D3 corpus collect expressions from one individual without any interaction.

Corpora in such static settings allow for the pursuit of interesting specific research-questions in a relatively simple domain. This is reflected for example in the organization and results of competitive events such as the TUNA-Challenge at ENLG 2009 (Gatt et al. (2009)). However, this type of setting differs significantly from actual human reference behaviour, where the collaborative aspect plays a central role. In fact, this point has been noted very early on (see for example Bolt (1980); Clark and Wilkes-Gibbs (1986); Heeman and Hirst (1995)).

As research on referring expressions has made progress, there has been an increasing move away from uniquely looking at the linguistic mode and towards studying the multimodal phenomenon of reference in domains approximating the complexity of actual human behaviour. Thus there has been ongoing work towards developing algorithms that deal with linguistic and various types of extra-linguistic information in the framework of one computational model (Kelleher et al. (2005); Byron et al. (2005); Kranstedt et al. (2006); Gergle and Kraut (2007); van der Sluis et al. (2008); Stoia et al. (2008); Prasov and Chai (2008)). This is partly a reflection of a general tendency towards the creation of multi-modal corpora in a number of domains (e.g. Qvarfordt et al. (2005); Schiel and Mögele (2008); Blache et al. (2009)).

Traditionally, many corpora have relied on manual work for their construction from speech and video recordings (e.g. Jokinen (2010)). In our corpus as well, linguistic information was transcribed and annotated manually, but extra-linguistic information was recorded and processed automatically. While reducing cost and increasing precision of the collected data, this approach might have disadvantages such as imposing limitations on the type of data that can be recorded (only automatically recordable). Overall, at the current stage of research both data collection by hand and automatic recording have their respective advantages.

Along with such progress in multi-modal data collection, recently corpora have been created in order to study referring expressions occurring in situated collaborative tasks. Given that such corpora seek to record increasingly realistic interactions and the role referring expressions play in them, there has been a large variety of research focal points that have been pursued: such as research into the co-occurrence of referring expressions with pointing behaviour (van der Sluis et al. (2008)), looking into the role of visual information in reference (Gergle and Kraut (2007)) and the influence of various pragmatic dimensions on the use of proximity markers (Byron and Stoia (2005)). One question that has not yet received significant attention is the interaction between the referring expressions and the actions performed by participants in collaboration tasks. However, in the context of working on a physical task, referring expressions are particularly important linguistic tools. Hence deepening our understanding of the interaction of linguistic information with extra-linguistic information on the actions in such a setting remains a critical task. The REX-J corpus, which we introduce in this paper, is a resource contributing to research in this area.

One major trend in research on human referring expressions has been based on the "centering theory" (Grosz et al. (1983); Brennan et al. (1987)). It employs rules and constraints that govern the relationship between discourse content and the linguistic choices made by the discourse participants, such as the choice of syntactic structure or the type of referring expression. In this theory, a discourse is fundamentally viewed as a dynamic process, where each sentence/utterance is a transition from an input state to an output state. Within this context, there have been studies for example attempting to predict which entities will be most salient at any given time (Poesio et al. (2000)). Based on the centering theory, there has also been a significant amount of work on Japanese referring expressions, more specifically on the interpretation of anaphora (Walker et al. (1994). Kameyama (1998)). We note that such work has not resulted in the construction of corpora per se.

This paper is organized as follows. Section 2 provides a succinct review of other existing multi-modal corpora of referring expressions in a collaborative task. Section 3 describes the construction of the REX-J corpus, including the annotation policies, and discusses the key characteristics of the collected data. Section 4 introduces two examples of usages of this corpus: an analysis of a specific type of referring expression as well as the generation of demonstrative pronouns in a collaborative task. Section 5 presents the conclusion, some potential research directions that might be pursued exploiting this corpus as well as our plans for future work.

2 Brief review of existing multi-modal corpora of referring expressions in collaboration

The construction of corpora in realistic domains is a critical task in order to achieve major progress in our understanding of referring expressions. A fundamental aspect that needs to be captured is the complexity which humans themselves are confronted with, when uttering or hearing referring expressions. Over the last decade, given the recognition that referring expressions are very frequently uttered in the context of a collaborative task (Clark and Wilkes-Gibbs (1986); Heeman and Hirst (1995)), a number of databases have been constructed in order to study referring expressions in such a domain. This tendency is also a reflection of the recognition that this area yields both a challenging, more realistic research domain, as well as promising possibilities for application, such as natural-language collaboration with robots on certain tasks (e.g. Foster et al. (2008); Kruijff et al. (2010)).

The COCONUT corpus (Di Eugenio et al. (2000)) is collected from keyboard-dialogs with enforced turn-taking and no interruptions allowed. The participants collaborate on a simple 2-D design task, i.e. buying and arranging furniture for two rooms. It resembles the TUNA corpus in tending to encourage very simple types of expressions by the participants. The COCONUT corpus has rich annotations at the linguistic and the intentional level including three kinds of features: (i) problem-solving utterance features, (ii) discourse utterance features and (iii) discourse entity features. However, it does not include extra-linguistic information, e.g. physical actions. Thus, in addition to the limited interaction naturalness, this corpus also has a limitation in terms of the recorded information, particularly extra-linguistic information. As an initial work in the construction of collaborative task corpora, the COCONUT corpus can be characterized as having a rather simple domain as well as a limited annotation. There has been some work utilizing this corpus for attribute selection as well as partner-specific adaptation (Jordan and Walker (2005); Gupta and Stent (2005)).

More recent work has mainly concentrated on seeking to overcome the shortcoming of domain complexity. Thus, the QUAKE corpus (Byron and Fosler-Lussier (2006)) – as well as the more recent SCARE corpus (Stoia et al. (2008)), which is an extension of QUAKE – is based on an interaction captured in a 3-D virtual reality world where two participants collaboratively carry out a treasure hunting task. Byron et al. (2005) exploits these two resources for research on the generation of noun phrases. While those two corpora deal with a rather complex domain (3-D virtual world), the participating subjects were only able to carry out limited kinds of actions (pushing buttons, picking up or dropping objects) relative to the complexity of the three-dimensional target domain. One of the reasons for this is that they focused on location-based references while we focused on event or action-based references.

As part of the JAST project, a Joint Construction Task (JCT) corpus was created based on two subjects participating in constructing a puzzle (Foster et al. (2008)). The setting of the experiment is quite realistic and natural in that both participants can equally intervene in the task. Based on this corpus, an analysis of a specific type of referring expression in task-oriented dialog is carried out. Furthermore, the role of eye-gaze in this setting is investigated (Bard et al. (2009)). While the authors note that the "transcribed speech was precisely time-aligned with all the visual and action components of the construction process", they do not provide further details on exactly what data they have recorded in their corpus. Despite the differences in some particulars of the task setting, further work on both the JCT as well as the REX-J corpus will allow interesting comparisons of specific phenomena between English and Japanese.

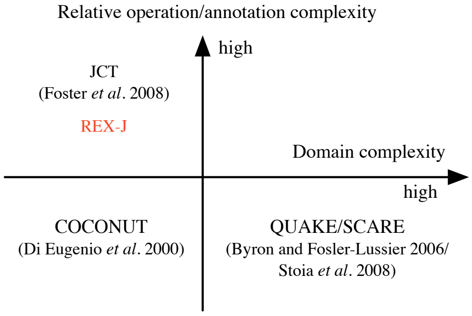

Relative operation/annotation complexity
high
JCT
(Foster *et al*. 2008)
REX-J
Domain complexity
high
COCONUT
(Di Eugenio *et al*. 2000)
QUAKE/SCARE
(Byron and Fosler-Lussier 2006/
Stoia *et al*. 2008)

Fig. 1 Schematic overview of recent multi-modal corpora

In contrast to these previous corpora, which were all in English, we created a Japanese corpus recording a whole range of information potentially relevant in the collaborative human reference process in a situated setting. “Situated dialog” in this context stands for a dialog where the participants are not solely communicating through linguistic means but share a space, and can act upon objects in that space. Participants had to solve the Tangram puzzle collaboratively. While the domain of our corpus is simple compared to the QUAKE and SCARE corpora, we allowed comparatively large flexibility in the actions necessary to achieve a task (such as flipping, turning and moving of puzzle pieces at different degrees), relative to the task complexity. Figure 1 provides a schematic comparison of the recent corpora discussed above in terms of the dimensions of relative operation/annotation complexity as well as domain complexity. Providing this relatively larger freedom of actions to the participants in the REX-J corpus together with recording the related information enables us to pursue research into new aspects of referring expressions that cannot be pursued based on previous corpora.

Among the existing task-oriented corpora, there are important distinctions in terms of the affordances of the experimental set-up of the task; that is whether pointing is possible (e.g. JCT corpus) or not (e.g. COCONUT corpus) and whether the dialog partners share an identical view of the task space (e.g. JCT corpus) or not (e.g. Map Task corpus (Anderson et al. (1991))). Obviously, the choices regarding these aspects strongly influence the character of the dialog and the types of referring expressions used by the participants. In addition, the number of speakers is an important distinction between corpora. While the corpora in static domains are mostly monolog such as in the TUNA corpus, situated corpora are built based on dialogs. Our corpus is also based on dialog. Furthermore, we chose a set-up where pointing was possible only for one participant while only the other could see the goal shape as well as the position of the mouse cursor. Thus, the two participants did not have an identical view of the task space.

Previous corpora have had a large bias towards English resources. While creating corpora in a language accessible world-wide is surely critical, the construction of language resources in other (and very different) languages is also a significant task that needs to be addressed by the community (Tokunaga et al. (2008)). Such work will also encourage the pursuit of comparative studies across languages of specific linguistic phenomena (e.g. the use of referring expressions) as well as the interaction between linguistic and extra-linguistic information in such a domain. Furthermore, in the study of referring expressions there has been notable interest over the recent period in the area of intersection of Computational Linguistics and Cognitive Science (van Deemter et al. (2009)). The REX-J corpus also makes a contribution in this broader sense.

In the construction of a corpus, a question to be considered is the trade-off between on the one side the “naturalness” of the domain and on the other the balance of the recorded data in various respects. Within the language analysis research community, the collection of “open domain” corpora – such as considered at the MUC and ACE-conferences (Grishman and Sundheim (1996); Strassel et al. (2008)) – has been the standard approach and there has not been a significant discussion on the need to collect balanced databases. On the other hand, within the language generation community there have been some arguments advanced underlining the need to create “balanced” corpora. Gatt et al. (2007) argue that ideally a corpus of referring expressions should both be balanced “semantically” (an equal number of situations requiring an equal set of attributes; a “balance of minimal descriptions”) and in terms of the target referents (different kinds of referents occur an equal number of times). van der Sluis et al. (2008) started work towards such a balanced multi-modal corpus. However, while in a static setting “balance” is a useful guideline for corpus construction, as van der Sluis et al. (2008) themselves note, in a collaborative domain this “arguably leads to a certain degree of artificiality in the conversational setting”. We note that our corpus is not

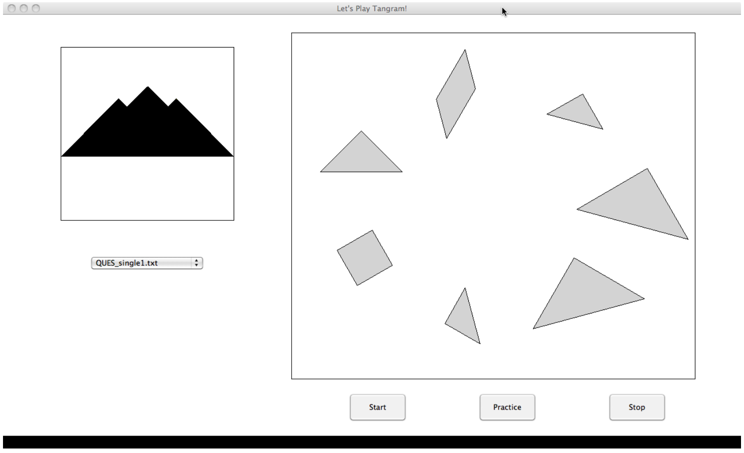


Fig. 2 Screenshot of the Tangram simulator

“balanced” as defined by Gatt et al. (2007). In a sense, we approach the question from the opposite end. We intend to collect a corpus enabling the study of various aspects of referring expressions through a comprehensive recording of the reference process in a collaborative task.

3 Construction of the corpus

3.1 The experimental set-up

In the process of building the REX-J corpus, we recruited 12 Japanese graduate students (4 females and 8 males) of the Cognitive Science department of Tokyo Institute of Technology, and split them into 6 pairs. All pairs knew each other previously and were of the same sex and approximately same age. [1] Each pair was instructed to solve the Tangram puzzle through collaborating with each other. The Tangram puzzle is a geometrical puzzle that originated in ancient China. The goal of this puzzle is to construct a given shape by arranging seven pieces of simple figures as shown in Figure 2. The pieces include two large triangles, a medium-size triangle, two small triangles, a parallelogram and a square.

With the aim of recording the precise position of every piece and every action the participants carried out during the solving process, we implemented a Tangram simulator in which the pieces on the computer display can be moved, rotated and flipped with simple mouse operations. The simulator displays two areas: a goal shape area (the left side of Figure 2) and a working area (the right side of Figure 2) where pieces are shown and can be manipulated.

We assigned a different role to each participant of a pair: a *solver* and an *operator*. Given a certain goal shape, the solver thinks of the necessary arrangement of the pieces and

[1] In Japanese culture, there exists the so-called *senpai-kouhai* relationship (relationship of senior to junior or socially higher to lower placed). Any different selection of experiment participants than we carried out would run the risk of including the effects of this social relationship and thus skew the collected data: such as the possible use of overly polite and indirect language, reluctance to correct mistakes etc. There has been some recent work on dealing with such cultural factors and creating standardized resources (Rehm et al. (2008)). In our study, we sought to avoid cultural effects as far as possible.

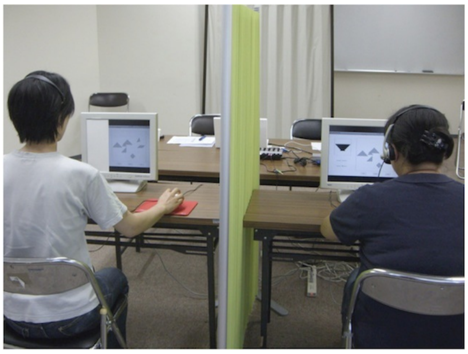


Fig. 3 Picture of the experiment setting

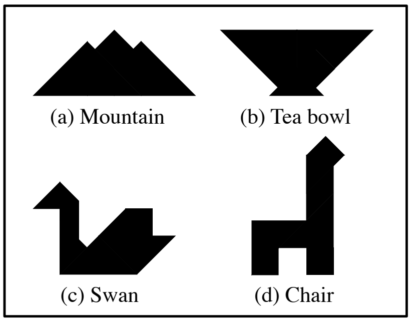


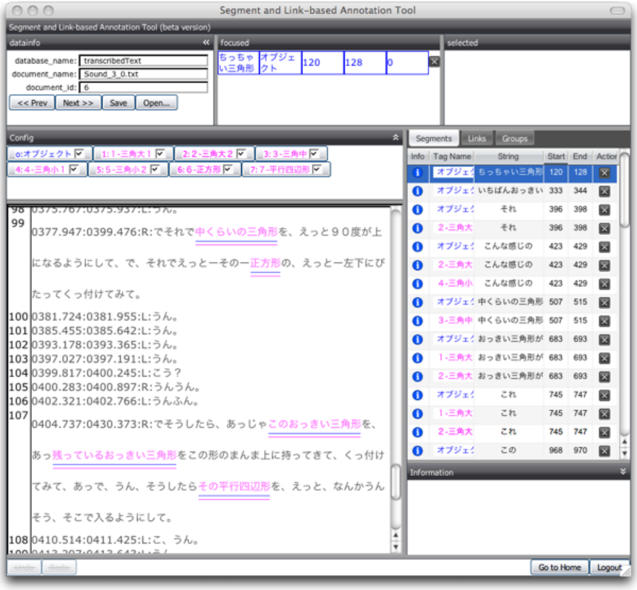
Fig. 4 The target shapes given to the subjects

gives instructions to the operator how to move them. The operator manipulates the pieces with the mouse according to the solver's instructions. In this interaction, we can expect relatively frequent uttering of referring expressions intended to distinguish certain pieces of the puzzle. In our Tangram simulator, all pieces are of the same color, thus color is useless in identifying a specific piece. Unlike the TUNA corpus, only size and shape are useful object-intrinsic attributes. Instead, we can expect other attributes such as spatial relations and deictic reference to be used more often in the dialogs.

The participants of a pair sit side by side as shown in Figure 3. A shield 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 only speech. Each participant has her/his own computer display sharing the working area where the movement of the mouse cursor and the pieces is shown in real-time. The operator (left of the separation) has a mouse for manipulation of pieces, but does not have a goal shape on the screen. In contrast, the solver (right of the separation) has a goal shape on the screen but not a mouse. Each participant pair is assigned 4 trials to each form two symmetric and two asymmetric shapes. The participants exchanged their roles after two trials, i.e. a participant first solves a symmetric and then an asymmetric puzzle as the solver and then does the same as the operator, and vice versa. The order of the puzzle trials is the same for all pairs. If we had a substantially larger pool of participants and the necessary resources, it might be of interest to vary the order of the tasks in order to counterbalance trials. However, because of insufficient number of participants, we maintained the same order throughout the data collection.

Figure 4 shows the four different goal shapes of the puzzle given to the subjects. In Cognitive Science, a wide variety of different kinds of puzzles have been employed extensively in the field of Insight Problem-Solving. This has been termed the "puzzle-problem approach" (Sternberg and Davidson (1996); Suzuki et al. (2001)) and in the case of physical puzzles has relatively often involved puzzle tasks of symmetric shapes (like the so-called T-puzzle, e.g. Kiyokawa and Nakazawa (2006)). In more recent work the Tangram puzzle has been used as a means to study various new aspects of human problem solving approaches, such as utilization of eye-tracking information (Baran et al. (2007)). In order to collect data as broadly as possible in this context, we set up puzzle-problems including both symmetrical ((a) and (b) of Figure 4) as well as asymmetrical ones ((c) and (d) of Figure 4).

Before starting the first trial as the operator, each participant had a short training exercise in order to learn the manipulation of pieces with the mouse. The initial arrangement of the pieces was randomized every time. We set a time limit of 15 minutes for the completion of one trial (i.e. construction of the goal shape). In order to prevent the solver from getting into deep thought and keeping silent, the 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. A trial ends when the goal shape is constructed or the time is up. Utterances by the participants are recorded separately in stereo through headset microphones in synchronization with the position of the pieces and the mouse operations. The positions of the mouse cursor and of all pieces as well as the mouse actions were automatically generated by the simulator. They were recorded in the pixel coordinates with time stamp at intervals of 1/65 second. We collected 24 dialogs (4 trials by 6 pairs) of about 4 hours and 17 minutes. The average length of a dialog was 10 minutes 43 seconds (with a standard deviation *SD* of 3 minutes 16 seconds).



database_name: transcribedText
document_name: Sound_3_0.txt
document_id: 6

ちっちゃい三角形	オブジェクト	120	128	0

☑ 0:オブジェクト ☑ 1:1-三角大1 ☑ 2:2-三角大2 ☑ 3:3-三角中
☑ 4:4-三角小1 ☑ 5:5-三角小2 ☑ 6:6-正方形 ☑ 7:7-平行四辺形

98 0375.767:0375.937:L:うん。
99 0377.947:0399.476:R:でそれで中くらいの三角形を、えっと９０度が上になるようにして、で、それでえっと一その一正方形の、えっと一左下にぴたってくっ付けてみて。
100 0381.724:0381.955:L:うん。
101 0385.455:0385.642:L:うん。
102 0393.178:0393.365:L:うん。
103 0397.027:0397.191:L:うん。
104 0399.817:0400.245:L:こう？
105 0400.283:0400.897:R:うんうん。
106 0402.321:0402.766:L:うんふん。
107 0404.737:0430.373:R:でそうしたら、あっじゃこのおっきい三角形を、あっ残っているおっきい三角形をこの形のまんま上に持ってきて、くっ付けてみて、あっで、うん、そうしたらその平行四辺形を、えっと、なんかうんそう、そこで入るようにして。
108 0410.514:0411.425:L:こ、うん。

Info	Tag Name	String	Start	End	Action
ℹ	オブジェ	ちっちゃい三角形	120	128	☒
ℹ	オブジェ	いちばんおっきい	333	344	☒
ℹ	オブジェ	それ	396	398	☒
ℹ	2-三角大	それ	396	398	☒
ℹ	オブジェ	こんな感じの	423	429	☒
ℹ	2-三角大	こんな感じの	423	429	☒
ℹ	4-三角小	こんな感じの	423	429	☒
ℹ	オブジェ	中くらいの三角形	507	515	☒
ℹ	3-三角中	中くらいの三角形	507	515	☒
ℹ	オブジェ	おっきい三角形が	683	693	☒
ℹ	1-三角大	おっきい三角形が	683	693	☒
ℹ	2-三角大	おっきい三角形が	683	693	☒
ℹ	オブジェ	これ	745	747	☒
ℹ	1-三角大	これ	745	747	☒
ℹ	2-三角大	これ	745	747	☒
ℹ	オブジェ	この	968	970	☒

Fig. 5 Screenshot of the SLAT annotation environment

3.2 Annotation

All 24 recorded dialogs were transcribed with a time code attached to each utterance. Since our central objective is the collection of referring expressions, we defined an utterance to be a complete sentence in order to prevent a referring expression being split across several utterances. In this respect, our "utterance" tends to be longer than those of ordinary dialog corpora in which an utterance is usually defined as a segment between pauses with a certain length.

In order to create an annotated corpus of referring expressions, we need to address two distinct steps in the annotation process. First, we have to determine *where* the referring expressions are (i.e. their span in a text) – here denoted as the expression identification step – and then identify their *referent* – the referent identification step. For both steps, we utilized the web-based multi-purpose annotation tool SLAT (Noguchi et al. (2008)). Figure 5 shows a screenshot of the SLAT annotation environment.

Two native Japanese annotators A_1 and A_2 (two of the authors) annotated 4 transcribed dialog texts independently, and resolved discrepancies by discussion. Based on the discussion, we decided on the following annotation guidelines.

- Only referring expressions whose referent is either one piece or a set of pieces in the working area are marked. Although there are other expressions referring to various types of referents such as a part of a piece and a location, representation of these kinds of referents is more difficult than that of a (set of) piece(s). Thus, we focus on expressions referring to the puzzle pieces in this corpus.
- A minimum span of a noun phrase with necessary information to identify a referent is marked. The span can include repairs with their reparandum and disfluency if any (Nakatani and Hirschberg (1993)). For example, in the case of an expression such as

  “*sakki zurasita migi no sankakkei iya heikousihenkei*
  ([the] triangle, no, [the] parallelogram on the right that you just moved)”,

  the information which initially modifies the mistaken object (only in Japanese) is included in the annotation. On the other hand, in an expression such as

  “*ôkii sankaku iya sikaku* ([the] big triangle, no, rectangle)”

  only “rectangle” is marked and “big” is not included since “big” modifies “triangle” but does not modify “square”. There are no rectangles of different sizes in the puzzle.
- Demonstrative adjectives are included in expressions, e.g. “*sono sankaku* (that triangle)”.
- Erroneous expressions are marked with a comment, e.g. “*tiisai sankaku* (small triangle)” referring to a square is marked with a comment “lexical error”.
- An expression without a definite referent (i.e. a group of possible referents or none) is marked with a prefix, followed by the sequence of possible referents, if there are any.
- All expressions appearing in muttering to oneself are excluded.

Based on these guidelines, the two annotators annotated the remaining 20 dialog texts independently. We evaluated inter-annotator agreement of these annotations by using the β-coefficient (Artstein and Poesio (2005)) for the expression identification, and the κ-coefficient for the referent identification. Although the κ-coefficient is a *de facto* standard in evaluating annotation reliability of corpora (Carletta (1996)), it is unweighted and thus does not differentiate between different degrees of disagreement. The fact that all disagreements are treated equally is a “serious limitation” in measuring discrepancies of spans of referring expressions (Artstein and Poesio (2008)). Hence the κ-coefficient is an inappropriate measure for annotator agreement on expression identification.

Expression identification In order to measure agreement on the identification of referring expressions, we need to account for a variety of degrees of disagreement, such as total mismatches or small boundary differences of one or two words. The β-coefficient has been used in measuring the inter-annotator agreement of multi-modal corpora, for example in Foster and Oberlander (2007). It has been noted that the actual value of β highly depends on how the annotation disagreements are categorized and how penalty weights are assigned to the categories. We adopted the same categories and penalty weights as Foster and Oberlander (2007) to calculate β. Other than the total match, disagreements in the annotation are categorized into three groups, given with their respective penalty weights in brackets: (a) strict subset ($1/3$), (b) overlap ($2/3$) and (c) absolute mismatch (1). Since we consider the matching based on the transcribed texts, this categorization can be done automatically.

The β-coefficient is calculated based on the ratio between the expected and observed disagreements in the annotation as follows:

$$\beta = 1 - \frac{D_{obs}}{D_{exp}} \tag{1}$$

where D_{obs} and D_{exp} respectively stand for the arithmetic mean of the observed disagreement values on the annotated spans, and the expected disagreement calculated from the distribution of annotated spans by each annotator. Values of β close to 1 reflect good agreement between the annotators, where the actual disagreement is much smaller in comparison to the expected disagreement. In contrast, values closer to 0 reflect bad agreement, where the observed disagreement approaches the expected disagreement.

The calculation of the observed disagreement is obvious. However, in terms of the expected disagreement, we need to note a difference to previous work on English corpora. The expected disagreement D_{exp} is calculated from the distribution of expression lengths. While in a language such as English, the length of an expression can be measured by the number of words, such a word count is not obvious in languages like Japanese (and many East Asian languages), which does not use explicit word delimiters. In fact, there is a significant amount of research on automatically detecting word boundaries in Japanese which has led to the development of a number of tools such as the morphological analyzer MeCab (Kudo et al. (2004)). This point indicates the type of issues faced with when building and evaluating non-English resources (see Tokunaga et al. (2008) for a broader discussion of this point in relation to Asian languages).

For the current purpose, we took a simplifying approach of disregarding words and measuring the length of expressions on a *character* basis. While this obliterates the distinction between semantic units (words) in the calculation and thus will have a tendency to slightly increase the expected disagreement (and hence increasing β-values), it provides a rough approximation. Computation according to the above yields $\beta = 0.664$. This means that the observed disagreement was less than half of the expected disagreement. We note that on the β-coefficient, no significance test is possible (Artstein and Poesio (2005)).

Referent identification We calculated the inter-annotator agreement of the reference identification using the cases categorized as "strict subset", "overlap" and "total match" in the previous step. As for the agreement of referent identification, we adopted a rigid agreement judgement, using the κ-coefficient for evaluation. Calculation yields $\kappa = 0.904$, which allows definite conclusions according to Krippendorf's scale to assess κ-values (Krippendorff (1980)).

Table 1 Inter-annotator agreement [%]

Annotator	Raw	i/Span	i/Ref
A_1	71.6	78.2	75.9
A_2	81.5	89.0	86.5

Cavicchio and Poesio (2009) discuss advantages and disadvantages of several recently employed statistical metrics for assessing the reliability of annotation of multi-modal corpora, clarifying there is not one "best" evaluation metric. As discussion and further work continues on developing more unified reliability measures, one way to deal with this problem at this point would be to measure inter-annotator agreement of a corpus in terms of several different metrics, thus providing evaluations from different standpoints. Based on such considerations, in Table 1 we show inter-annotator agreement in percentages of both annotation steps discussed in this section. The row for A_1 shows the agreement values considering A_2's annotation as a gold standard, and likewise respectively for the row for A_2. "Raw" here denotes the percentage of the cases where one annotator's annotation is exactly the same as the other's, i.e. both recognized the same span of a text as a referring expression and identified the same referent(s). "i/Span" and "i/Ref" respectively denote the rate of agreement when ignoring discrepancy in span and referent(s) respectively.

Table 2 Number of expressions by Dialog

Dialog ID	Solver	Operator	Total
D_0	37	5	42
D_1	44	5	49
D_2	22	4	26
D_3	14	1	15
D_4	72	23	95
D_5	55	5	60
D_6	55	3	58
D_7	53	4	57
D_8	74	28	102
D_9	43	3	46
D_10	11	2	13
D_11	67	29	96
D_12	44	8	52
D_13	24	1	25
D_14	38	18	56
D_15	16	3	19
D_16	39	5	44
D_17	45	4	49
D_18	35	0	35
D_19	54	1	55
D_20	61	7	68
D_21	57	7	64
D_22	47	0	47
D_23	69	0	69
Total	1,076	166	1,242
Aver.	44.8	6.9	51.8
SD	18.1	8.5	23.8

Table 3 Number of expressions by Speaker

Speaker ID	Solver	Operator	Total
S_0	81	5	86
S_1	36	10	46
S_2	127	7	134
S_3	108	28	136
S_4	117	31	148
S_5	78	31	109
S_6	68	21	89
S_7	54	9	63
S_8	84	1	85
S_9	89	9	98
S_10	118	0	118
S_11	116	14	130
Total	1,076	166	1,242
Aver.	89.7	13.8	103.5
SD	28.3	11.2	31.3

3.3 The recorded data

We collected a total of 1,443 tokens and 425 types (different surface realizations) of referring expressions in 24 dialogs. The size of our corpus is roughly comparable to both the SCARE corpus (15 dialogs with 1,700 expressions) and the COCONUT corpus (24 dialogs with 1,100 utterances). Our asymmetric experimental setting tended to encourage referring expressions from the solver, while the operator mainly employed referring expressions in order to confirm his understanding of the solver's instructions. This is reflected in the number of referring expressions by the solver (1,243 tokens) largely outnumbering those of the operator (200 tokens).

Table 4 Syntactic and semantic features of identified referring expressions

	Feature	Label	Tokens	Example
(a)	*demonstrative*		742	
	adjective	dad	194	"*ano migigawa no sankakkei* (that triangle on the right side)"
	pronoun	dpr	548	"*kore* (this)"
(b)	*attribute*		795	
	size	siz	223	"*tittyai sankakkei* (the small triangle)"
	shape	typ	566	"*ôkii sankakkei* (the large triangle)"
	direction	dir	6	"*ano sita muiteru dekai sankakkei* (that large triangle facing to the bottom)"
(c)	*spatial relations*		147	
	projective	prj	143	"*hidari no okkii sankakkei* (the big triangle on the left)"
	topological	tpl	2	"*ôkii hanareteiru yatu* (the big distant one)"
	overlapping	ovl	2	"*sono sita ni aru sankakkei* (the triangle underneath it)"
(d)	*action-mentioning*	act	85	"*migi ue ni doketa sankakkei* (the triangle you put away to the top right)"

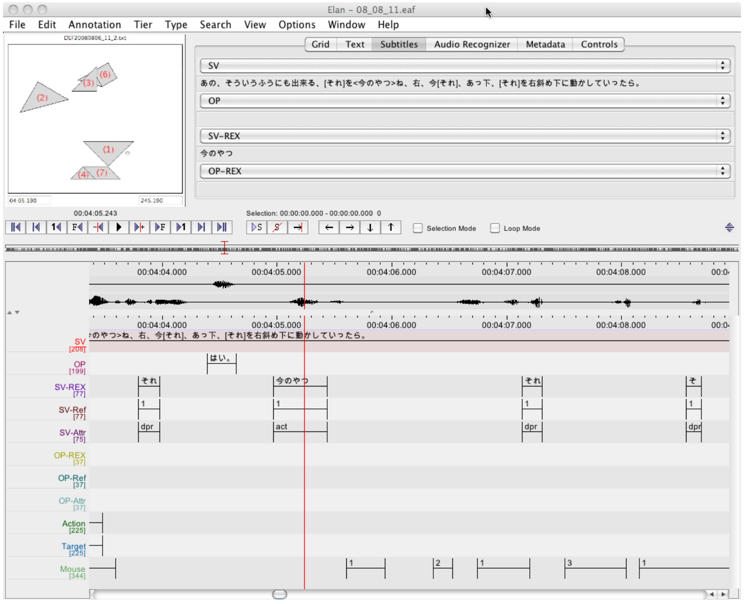
The annotated referring expressions include indefinite expressions[2] whose referents cannot be determined uniquely. For example, since there are two large triangles in the puzzle, there might be a case where the expression "*ôkii sankaku* (a large triangle)" can refer to either one of them. We annotated these indefinite expressions with an "indefinite" attribute together with possible referent candidates. We exclude such indefinite expressions as well as expressions explicitly referring to multiple pieces from the further analysis reported hereafter. The number of the excluded expressions is 201, about 15% of the total. Thus, the number of the remaining expressions under consideration which refer to a single piece is 1,242. Table 2 and Table 3 further detail the expressions collected by dialog and by speaker.

One of the annotators further annotated these expressions with the syntactic and semantic features listed in Table 4. The following syntactic/semantic features in referring expressions can be identified: (a) demonstratives (adjectives and pronouns), (b) intrinsic attributes of pieces, (c) spatial relations and (d) actions on an object. Note that multiple features can be used in a single expression. The right-most column shows an example from the corpus with its English translation. The identified feature in the referring expression is underlined.

We utilized the multi-modal annotation tool ELAN [3] in order to merge the manually annotated linguistic information and the extra-linguistic information generated by the Tangram simulator. The extra-linguistic information includes (i) the action on a piece, (ii) the coordinates of the mouse cursor and (iii) the position of each piece in the working area. The actions and the mouse cursor positions are recorded at intervals of 1/65 second. This information is then automatically abstracted into (i) a time span labeled with an action symbol ("move", "rotate" or "flip") and its target object number (1-7), and (ii) a time span labeled with a piece number which is under the mouse cursor during that span. The position of puzzle pieces is updated and recorded with a time stamp whenever the position of any piece changes. We did not merge the information about piece positions into the ELAN files and kept it in separate files. As a result, we have 11 time-aligned ELAN Tiers. They are repre-

[2] Since Japanese has no article marking definiteness, distinguishing between definite and indefinite expressions depends on their contexts.

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



Elan – 08_08_11.eaf

SV

あの、そういうふうにも出来る、[それ]を<今のやつ>ね、右、今[それ]、あっ下、[それ]を右斜め下に動かしていったら。

OP

SV–REX

今のやつ

OP–REX

Selection: 00:00:00.000 - 00:00:00.000 0

Fig. 6 Snapshot of the REX-J corpus in ELAN

sented by the lines marked by colored labels in the bottom half of Figure 6. We have Tiers such as “SV-REX” (referring expressions by the solver) and “Mouse” (the piece on which the mouse is hovering).

4 Two example uses of the REX-J corpus

Multi-modal corpora of referring expressions have been utilized to pursue research in a range of fields, often with a view towards language generation. Jordan and Walker (2005) utilized the COCONUT corpus and worked on the generation of content descriptions (attribute sets) of referring expressions. Stoia et al. (2006) sought to generate referring expressions given NP frame slots as input, utilizing various context variables as features in a machine-learning approach. In this section, we provide results from two investigations, one each in the area of language analysis and generation, pointing to some of the various potential uses of the REX-J corpus.

4.1 Action-mentioning expressions

Analyzing the expressions that appear in our corpus, we found a type of referring expression that mentions an action on an object (Spanger et al. (2009a)), such as “*ima hanten sita yatu* (the one you just flipped over)”. This type of referring expression – which we call an *action-mentioning expression* (“AME” hereafter) – has attracted less attention in previous work in the context of collaborative interaction. AMEs are related in a broad sense to English expressions such as taken up in Novak (1986), who sought to generate a description of a traffic scene using events (e.g. “taking over”, etc). Novak (1986) concentrated on event

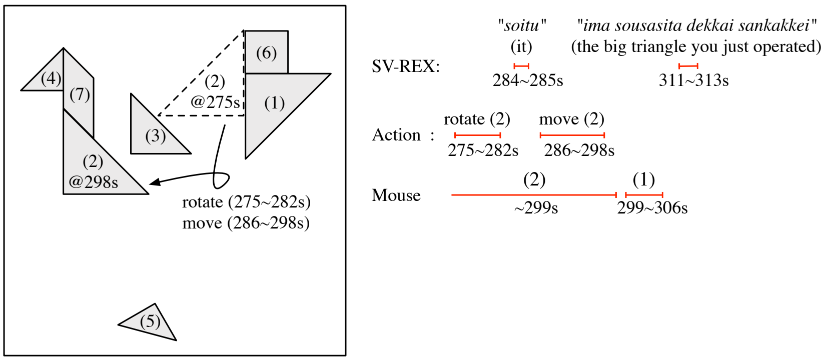


Fig. 7 Example of an AME, showing recent movement mouse and referent

verbalization (such as “the car is moving”) employed in order to distinguish an object in a description task. He provided no further analysis of the relation of this kind of expression with other features in referring expressions. The domain he deals with is a simple observed situation not including any collaboration and thus very different from the domain discussed here. Furthermore, we also note that AMEs are different from *haptic-ostensive* referring expressions discussed in Foster et al. (2008) since AMEs are not necessarily accompanied by a simultaneous physical operation on an object.

In contrast to Novak (1986), we focus on actions carried out on objects in a collaborative task setting. Considering this phenomenon in the context of a situated dialog allows us to investigate in detail the interaction of actions on objects by participants and the use of AMEs. Figure 7 shows an example of an AME from the corpus (Dialog 21) with its context. On the right side of the figure, we show the time-aligned information on the referring expressions (SV-REX), operations on pieces (Action) as well as the time periods when the mouse is over a certain piece (Mouse). The referent of the AME in this example is the recently moved large triangle (2), with its previous position indicated in broken lines. Generally, in a collaborative task as considered here, we can assume actions to be very salient for both participants. In comparison to other shorter expressions such as demonstratives or the simple use of the shape of the piece, an AME might be longer and thus has less ambiguity.

On average, every pair used about 14 AMEs over the four trials. Although there was variation in the usage of AMEs among the pairs ($SD \approx 9$), all 6 pairs of participants used AMEs, indicating that it is a fundamental type of expression for this task setting. 84 out of 85 AMEs were used by the solvers. This is explained by the asymmetric setting of our experiment.

In order to resolve AMEs, factors such as the saliency of actions mentioned in the expressions as well as *recency* will be useful. Traditionally, recency means that the more recently an object was mentioned, the more salient it will be, thus more likely to be the referent of a referring expression. We note that in other domains, specific actions might be explicitly referred to (e.g. “please take the pink dress I gave you for your birthday last year”), without requiring any notion of recency. However, in the context of a collaborative task domain, such mention of actions taken a long time ago will be rare and rather exceptional. It has been pointed out in previous work that recency is also one of the most significant factors for the resolution of demonstrative pronouns (Mitkov (2002)). As shown in Table 4, demonstrative pronouns along with the attribute “shape” are the most frequent feature of referring expressions in the REX-J corpus. More generally, demonstrative pronouns such as “*kore* (this)”, "*sore* (it)" and "*are* (that)" are a fundamental device for referring to objects (Halliday and Hassan (1976)).

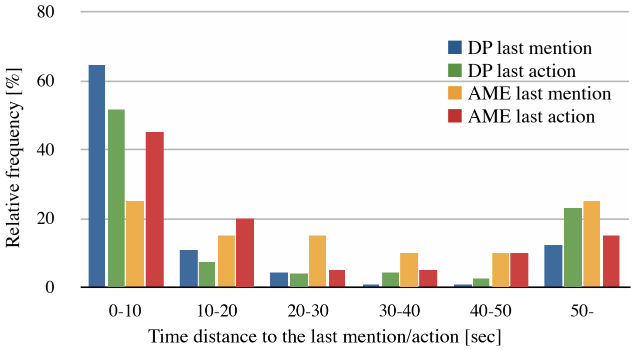


Fig. 8 Relative frequencies of AMEs and DPs over the time distance to the last mention/action

Based on the extensive previous research underlining the importance of recency for the resolution of demonstrative pronouns and our intuition that recency will be also useful for the understanding of AMEs, we looked into the relationship between the frequency of AMEs/demonstrative pronouns (DPs) and the time distance to both the last linguistic mention and the last physical action with regard to the referent. Since AMEs were used almost exclusively by the solvers, we considered only DPs used by the solvers for comparison.

Defining recency in terms of the last linguistic mention of an object corresponds to the traditional concept of textual recency as employed in reference resolution. As shown in Figure 8, we confirmed the very fast decay of the use of DPs over the first 30 seconds from the last mention and from the last action. Relative frequencies for those two cases seem to show a similar decay pattern at first glance. However, a χ^2 test on the absolute frequencies of "DP last mention" and "DP last action" over the time distance showed a significant difference at 0.01% level. This difference suggests that recency in terms of the last action is also useful for reference resolution (Iida et al. (2010)).

Figure 8 also indicates that the relative frequencies of AMEs over the time distance to the last action show a quite strong decay pattern similar to DPs. In contrast, the relative frequencies of AMEs when measured over the time distance to the last mention decrease much slower. We confirmed a significant difference between the decay patterns of "AME last mention" and "AME last action" by a χ^2 test at 5% level.

The analysis of the frequency decay of DPs and AMEs over the time distance provides new perspectives on the concept of recency. In particular, previous research considered recency mainly in a textual context focusing almost exclusively on linguistic information (Walker et al. (1994)). Hence, recency was often measured by metrics such as the number of sentences between a referring expression and its antecedent. However, our analysis indicates that the recency of the last linguistic mention shows a different tendency from the recency of the last physical action. There is a possibility that the action-based recency can contribute to resolving AMEs and DPs. Thus, we propose an extended notion of recency in a physical sense, i.e. how far back an object was physically manipulated.

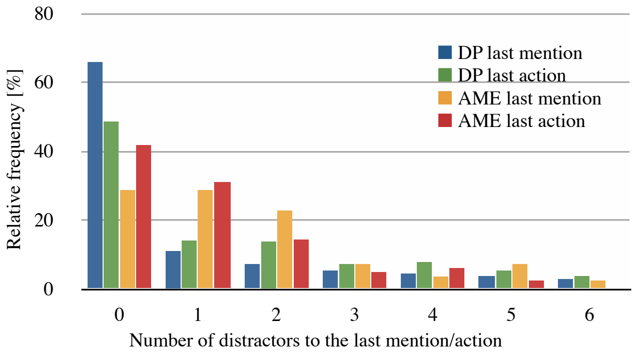


Fig. 9 Relative frequencies of AMEs and DPs over the number of distractors to the last action/mention

In addition to measuring recency in terms of time distance, we also consider a different metric of recency in terms of the number of distractors. Here ”distractors” are defined as either those objects manipulated since the last action on the referent or those objects mentioned since the last mention of the referent. Figure 9 shows relative frequencies of DPs and AMEs based on this metric.

As in Figure 8, the decay of relative frequencies of DPs over the number of distractors when measured from the last mention and from the last action, seems to show a similar tendency in Figure 9. However, a χ^2 test revealed a statistically significant difference (at 5% level) between “DP last mention” and “DP last action”. This result is similar to the result of our analysis of frequencies measured over the time distance. For AMEs, “AME last action” seems to decrease faster than “AME last mention”. However, a χ^2 test did not show a significant difference between them.

We examined two factors for recency: the metric for measuring distance (the time distance vs. the number of distractors) and the definition of the starting point of intervals (the last mention vs. the last action). Overall, for both DPs and AMEs, there were different tendencies of relative frequencies when measured in relation to the last action or the last mention. In contrast, the results showed similar tendencies for both distance measuring metrics. This suggests the need to investigate any possible difference in impact of those factors on the resolution of DPs and AMEs.

Table 5 Classification of action-mentioning expressions

Category	Tokens	Ratio [%]
(a) Temporal adverbials w/ verb	55	65
(b) Temporal adverbials w/o verb	22	26
(c) Verb only	8	9
Total	85	100

AMEs consist of a noun modified by an adnominal phrase, which often includes a verb describing an action and temporal adverbials. Considering the surface structure of AMEs, they can be divided into three categories based on the elements constituting the adnominal phrase: (a) combination of a temporal adverbial with a verb like “*sakki kaiten saseta ôkii* *sankaku* (the large triangle you just turned around)”, (b) a temporal adverbial without a verb, i.e. verb ellipsis such as “*ima no heikousihenkei* (the parallelogram you are [verb-ing] right now)” and (c) a verb without a temporal adverbial such as “*sono itiban ue ni oita sankakkei* (that triangle [you] put at the very top)”. Table 5 shows the frequency distribution of these three categories in the corpus.

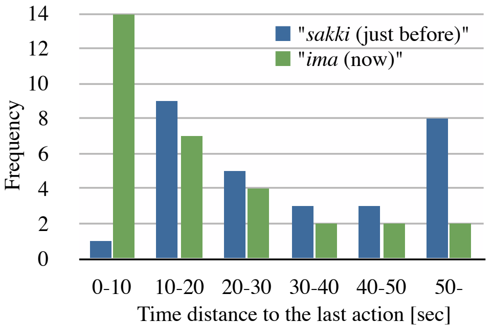


Fig. 10 Frequencies of temporal adverbials in AMEs over the time distance to the last action on the referent

The second category, which includes verb ellipsis, would be rare in English, but it is quite natural in Japanese. This is also related to differences in syntax between Japanese and English; while in English expressions such as “now” are temporal adverbials and as such modify a verb, Japanese temporal adverbials can be adnominals with the particle “*no* (of)”. In addition, comparing the numbers of the second and the third categories, it is interesting to see that temporal adverbials tend to be more explicitly mentioned. This kind of consideration is an example of the comparative analysis fostered by the creation of resources in languages other than English.

The dominant temporal adverbials used by the participants were “*sakki* (just before)” and “*ima* (now)“, e.g. *sakki no NP* (the NP [*verb*-ed] just before)” or “*ima no NP* (the current NP/the NP [you are *verb*-ing] now/the NP [*verb*-ed] just before)”. “*Ima*” generally refers to the current time point (“now”). It can, however, refer to the recent past as well and thus is ambiguous. We note that in case a verb is explicitly mentioned, disambiguation is possible based on its aspect (progressive vs. perfect). Participants tended to use “*ima*” largely in its perfect meaning (completed action). The distribution of use of “*ima*” in its perfect meaning in comparison to its progressive meaning was about 2:1 (31 instances/15 instances). The distribution of the two types of temporal adverbials “*sakki*” and “*ima*” was about 2:3 (29/46). The higher frequency of “*ima*” might be explained by its dual meanings (progressive and perfect) in contrast to the exclusive use of “*sakki*” for past actions.

Figure 10 shows the distribution of “*sakki* (just before)” and that of “*ima* (now)” in its perfect meaning, over the time distance to the mentioned action. For actions occurring within a time frame of about 10 seconds previous to an expression, participants had an overwhelming preference for “*ima*”, while for actions further in the past, participants preferred “*sakki*”. We carried out a χ^2 test over the differences between “*sakki*” and “*ima*”, which showed a significant difference at 1% level.

In the above discussion, we saw that the use of temporal adverbials in AMEs is highly dependent on the temporal distance to the action. Considering this phenomenon, we note an interesting parallel with demonstrative pronouns, whose use is very much tied to the concept of spatial distance (i.e. the distinction between *kore* (this) and *sore* (that)).

In our analysis of AMEs as shown in Figures 8 and 9, we discussed how recency can be useful in terms of a physical action as well as a linguistic mention. We furthermore showed how the distance to these events can be defined in different ways: temporal intervals vs. the number of distractors. The comparison of AMEs with pronouns led us to extend the traditional definition of recency in two directions, both emphasizing the importance of physical actions.

4.2 Generation of Demonstrative Pronouns

Pursuing research into new aspects in the generation of referring expressions is a further possible use of the REX-J corpus. In recent work both Jordan and Walker (2005) and Stoia et al. (2006) studied the generation of referring expressions in specific dialog settings, with Stoia et al. (2006) particularly stressing the importance of integrating extra-linguistic information as part of the context. Thus, they introduced “spatial/visual features” in addition to dialog history. Jordan and Walker (2005) provide a number of theoretically motivated features such as “intentional influences” (modeling the current state of the task as well as agreement by the participants). The REX-J corpus allows us to reflect on the role of extra-linguistic information in the production of referring expressions (Spanger et al. (2009b)). This section briefly describes our recent work on the generation of demonstrative pronouns in a collaborative task.

In fact, while accounting for certain types of extra-linguistic information, the above approaches neither dealt with information on current operations nor on the actions that have been performed by participants during the course of the collaboration. In previous work, it has been pointed out that in a collaborative task (e.g. the construction of an object), participants’ actions on objects impact their reference behaviour to a significant degree (Foster et al. (2008)). Thus, we can project that the development of a model integrating information from the linguistic as well as other modes will be vital to deepening our understanding of referring expressions in this domain.

Seeking to replicate human use of demonstrative pronouns in the corpus, we employed a machine learning approach utilizing linguistic as well as extra-linguistic information. This information is then employed as features in a Support Vector Machine (SVM) – a supervised learning method for binary classification (Vapnik (1998)).

Table 6 Features for generating demonstrative pronouns

Dialog History	Action History	Current Operation
D1: time distance to the last mention of the target	**A1**: time distance to the last action on the target	**O1**: the target is under operation
D2: last expression type referring to the target	**A2**: last operation type on the target	**O2**: the target is under the mouse
D3: number of other pieces mentioned during the time period of D1	**A3**: number of other pieces that were operated during time period of A1	
D4: time distance to the last mention of another piece	**A4**: time distance to the last operation on another piece	
D5: the target is the last mentioned piece	**A5**: the target is the latest operated piece	

Table 7 Learnt weight of features

Rank	Feature	Weight	Rank	Feature	Weight
1	O2 = yes	1.696	15	A3	0.005
2	D5 = yes	0.309	16	A2 = rotate	0.001
3	D4 $\leq$ 10	0.262	17	A2 = move	0.000
4	O1 = yes	0.229	18	D1 $\leq$ 20	-0.006
5	D2 = pron	0.124	19	A1 $\leq$ 20	-0.023
6	A1 $\leq$ 10	0.108	20	A1 $>$ 20	-0.037
7	D1 $\leq$ 10	0.102	21	D4 $>$ 20	-0.057
8	D2 = other	0.098	22	A4 $\leq$ 10	-0.097
9	A5 = no	0.069	23	O1 = no	-0.118
10	A4 $>$ 20	0.066	24	O2 = no	-0.147
11	A4 $\leq$ 20	0.060	25	D5 = no	-0.151
12	D3	0.060	26	D1 $>$ 20	-0.153
13	A2 = flip	0.038	27	D4 $\leq$ 20	-0.356
14	A5 = yes	0.010			

We defined the 12 features shown in Table 6 in order to represent a specific situation of the task context, at the time point when a target is mentioned. The features are categorized into three categories: dialog history features (D1$\sim$D5), action history features (A1$\sim$A5) and the current operation features (O1 and O2). While the features D1, A1 and D4, A4 measure time distances, the features D3 and A3 measure the number of other pieces referred to/acted upon. We recall that overall decay patterns of relative frequencies measured by those two metrics were very similar (see Figure 8 and 9). However, this does not necessarily mean both metrics have an equal impact on the generation of DPs. The addition of features A3 and D3 will allow us to look into this question, in comparison to the time distance metric encoded in features D1, A1 and D4, A4.

The dialog features model the dialog history, while the operation and action features capture key aspects of the physical actions that might have an impact on the accessibility of the target and thus the usage of demonstratives.

In order to provide a more detailed view of the data encoded, all features were then split up according to their respective values. The time distance features (A1, A4, D1 and D4) were discretized into three intervals: $(0,10]$ sec, $(10,20]$ sec and $(20,\infty]$ secs based on the analysis of action-mentioning expressions in the previous section, particularly Figure 8. This results in a total of 27 features as shown in Table 7.

Our aim is to automatically decide whether or not to use a demonstrative pronoun to mention a target in a specific situation which is described in terms of the above features. We constructed an SVM classifier which classifies a pair comprised of a target piece and a situation represented by the above features into two classes: "demonstrative pronoun" and "other". We employed the SVM-light software (Joachims (1998)) with 1,242 instances of referring expressions referring to one target object, extracted from our corpus and applied a 10-fold cross validation.

Table 8 Results of classification

Features	Recall	Precision	F-measure
Baseline	0.653	0.656	0.654
All	0.811	0.664	0.730
w/o D1$\sim$D5	0.822	0.652	0.727
w/o A1$\sim$A5	0.768	0.685	0.724
w/o O1, O2	0.585	0.576	0.580

Table 8 shows the overall results of the classification, with a separate row each for removing features of a category (dialog history, action history and current operation) from the set of all features, here denoted by “All”. The baseline suggests the use of a demonstrative pronoun, whenever the most recently mentioned object is the target, and suggests the non-use of a demonstrative pronoun otherwise. We see that the baseline achieves a smaller F-measure than when using all features. In contrast, removing the operation features leads to a worse F-measure than the baseline.

While removing either dialog or action history (the rows “w/o D1∼D5” and “w/o A1∼A5”) has a negligible effect on the overall F-measure, we note significant performance deterioration when removing the current operation features (O1 and O2). This reflects the fact that information of the ongoing action has a particularly strong impact on the use of demonstrative pronouns.

The asymmetric setting of the REX-J corpus results in most referring expressions being used by solvers (e.g. out of all demonstrative pronouns, 401 are by the solver, 147 by the operator), who are not allowed to point at pieces. In a situation where the mouse cursor is on the target, we are in fact dealing with a *joint action*, with the solver using a linguistic expression while the operator points to a piece. Differentiating it from a traditional pointing action, we might be able to call this phenomenon “collaborative pointing”. This is closely related to the joint attention effect discussed in Diessel (2006).

The results discussed here – in spite of the language difference of Japanese vs. English – also support the claim of Piwek (2007), that speakers tend to utter shorter linguistic expressions when using pointing actions in a similar setting to ours (in their symmetric setting, however, both participants were able to point at objects). This had also been noted much earlier by Brennan and Clark (1996). In fact, deictic use of pronouns is dominant in our corpus; 402 demonstrative pronouns out of 548 were used with the mouse cursor being on the target.

Table 7 shows the ranked list of the learnt weights of features which were calculated by using a linear kernel and all training instances. The weight of a feature reflects the importance of the respective feature in classification. We note that the two top-weighted features encode information on the current operation (O2) as well as the dialog history (D5). This confirms the point mentioned at the outset; namely the need to integrate linguistic and extra-linguistic information into a unified account.

The relatively high rank of “D2=pron” could be interpreted in such a way that a piece mentioned by a pronoun the last time tends to be subsequently mentioned by pronoun. This observation is consistent with past research on anaphora resolution (Mitkov (2002)).

Another remarkable tendency can be seen in the ranks of features A1 and A4. Among A1 features, the most recent one (≤ 10) has the highest rank (6), while the two more distant ones (≤ 20, > 20) have lower ranks (19 and 20). In contrast, the ranks of A4 features show the opposite tendency. That is, the most recent one (≤ 10) has the lowest rank (22), while the two more distant ones (> 20, ≤ 20) have higher ranks (10 and 11). This indicates that in order to use pronouns, the target is better to have been operated recently (high rank of A1 ≤ 10). In contrast, the other pieces are better to have been operated a long time ago (higher rank of A4 ≤ 20). Interestingly, there is no such clear tendency for their counterparts among the dialog history features, D1 and D4. In addition, features D3 and A3 reside in close ranks (12 and 15); this means the number of other pieces operated/mentioned during the period from the last mention/operation of the target to the referring expression has a similar effect. We see that 5 out of 12 time distance-based features are higher ranked than either D3 and A3. This indicates that rather than distractor-based features, time based-features are more effective for the generation of DPs.

Table 9 Performance of individual feature combinations

ID	Feature	Prec.	Recall	F
1	O2=yes	0.305	1.000	0.468
2	D5=yes	0.635	0.817	0.715
3	D4 $\leq$ 10	0.720	0.709	0.715
4	O1=yes	0.720	0.709	0.715
5	D2=pron	0.720	0.709	0.715
6	A1 $\leq$ 10	0.720	0.709	0.715
7	D1 $\leq$ 10	0.720	0.709	0.715
8	D2=other	0.720	0.709	0.715
9	A5 = no	0.720	0.709	0.715
10	A4 $>$ 20	0.720	0.709	0.715
11	A4 $\leq$ 20	0.635	0.817	0.715
12	D3	0.720	0.709	0.715
13	A2=flip	0.720	0.709	0.715
14	A5=yes	0.720	0.709	0.715
15	A3	0.720	0.709	0.715
16	A2=rotate	0.720	0.709	0.715
17	A2=move	0.720	0.709	0.715
18	D1 $\leq$ 20	0.720	0.709	0.715
19	A1 $\leq$ 20	0.652	0.822	0.727
20	A1 $>$ 20	0.652	0.822	0.727
21	D4 $>$ 20	0.652	0.822	0.727
22	A4 $\leq$ 10	0.652	0.822	0.727
23	O1=no	0.652	0.822	0.727
24	O2=no	0.652	0.822	0.727
25	D5 = no	0.652	0.822	0.727
26	D1 $>$ 20	0.652	0.822	0.727
27	D4 $\leq$ 20	0.664	0.811	0.730

Given these feature weights, we investigated the impact of each feature by evaluating the performance of feature combinations which were generated by adding one feature at a time in descending order of their weight, i.e. a feature combination K includes feature 1 through feature K. Table 9 shows the development of precision, recall and F-measure over the feature combinations 1–27. Following the first two feature combinations, the F-measure grows only very slightly and only at two more places: where features 19 (A1 $\leq$ 20) and 27 (D4 $>$ 20) are added. Interestingly, we note that for feature combinations 2 and 3 the F-measure remains the same, while the precision and recall values are different.

5 Conclusions and future work

This paper presented the REX-J corpus, which can be used as a resource in order to pursue research on referring expressions occurring in a collaborative task. This corpus captures linguistic information in synchronization with information on the actions carried out by the participants to solve the Tangram puzzle. Through outlining the construction of the corpus, the annotation scheme as well as the collected data, we discussed some of the particularities of Japanese in comparison with English, and how they impact the creation and use of this resource. To show the potential of the corpus, we provided two examples of research on the analysis and generation of referring expression by using this corpus[4].

This corpus can also be a valuable contribution with a view towards stimulating broader research at the intersection of Cognitive Science and Computational Linguistics, since it allows us to study recorded linguistic data in combination with the actions having occurred as well as the current state of the ongoing collaboration. In addition to research from the Computational Linguistics perspective outlined in the previous section, work based on the REX-J corpus is currently being pursued from a Cognitive Science perspective (Kuriyama et al. (2009)) as well. More broadly, integrating recent work in Cognitive Science on problem solving with an analysis of the referring expressions would help to address questions such as what the impact of a specific task or a state of task is on the use of different types of referring expressions.

[4] The REX-J corpus will be distributed through GSK (Language Resources Association in Japan; http://www.gsk.or.jp/index_e.html).

In addition, as a non-English linguistic resource, this corpus can contribute to a broadening of research on referring expressions as well as to the development of language-universal models.

Utilizing the REX-J corpus, time-aligned information on linguistic interaction and actions can be analyzed in order to develop more general models for referring expressions by integrating linguistic and extra-linguistic information. Viethen and Dale (2008) have discussed individual differences in referring expressions in a static setting. As any real application would have to deal with this in a dynamic setting, looking into individual differences within this corpus would be an interesting research direction. Furthermore, as this corpus captures an intensive collaboration on a task by two subjects, questions of alignment that have recently received attention (Janarthanam and Lemon (2009); Buschmeier et al. (2009)) can be addressed in a collaborative setting. Although we focused on expressions referring to a single object, generally reference to a group of objects is also an important linguistic device. We have previously discussed expressions referring to a group of objects in a static setting (Funakoshi and Tokunaga (2006)). The REX-J corpus not only allows for the pursuit of research analyzing such expressions in a dynamic setting, but also for research on how they interact with the actions by the participants.

In future work, we plan to collect a parallel corpus in English based on the same (or a very similar) task setting, laying the basis for further comparative research of phenomena found in the Japanese corpus (Tokunaga et al. (2010)). Furthermore, the current setting still excludes many modes of extra-linguistic information that are normally available in a real-world environment, such as information of eye-gaze. We intend to extend the types of data recorded in the current corpus to other modes in order to further approach a real-world environment.

References

Anderson, A. H., M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weiniert: 1991, ‘The HCRC Map Task Corpus’. *Language and Speech* **34**(4), 351–366.

Artstein, R. and M. Poesio: 2005, ‘Kappa^3 = Alpha (or Beta)’. Technical Report CSM-437, University of Essex.

Artstein, R. and M. Poesio: 2008, ‘Inter-Coder Agreement for Computational Linguistics’. *Computational Linguistics* **34**(4), 555–596.

Baran, B., B. Dogusoy, and K. 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*. pp. 555–563.

Bard, E. G., R. Hill, M. Arai, and M. E. Foster: 2009, ‘Accessibility and attention in situated dialogue: Roles and regulations’. In: *Proceedings of the Workshop on Production of Referring Expressions Pre-CogSci 2009*.

Blache, P., R. Bertrand, and G. Ferré: 2009, ‘Creating and exploiting multimodal annotated corpora: the ToMA project’. In: M. Kipp, J.-C. Martin, P. Paggio, and D. Heylen (eds.): *Multimodal Corpora*. Springer-Verlag, pp. 38–53.

Bolt, R. A.: 1980, ‘“Put-that-there”: Voice and gesture at the graphics interface’. In: *Proceedings of the 7th annual conference on Computer graphics and interactive techniques (SIGRAPH 1980)*. pp. 262 – 270, ACM.

Brennan, S. E. and H. H. Clark: 1996, 'Conceptual Pacts and Lexical Choice in Conversation'. *Journal of Experimental Psychology: Learning, Memory and Cognition* **22**(6), 1482–1493.

Brennan, S. E., M. W. Friedman, and C. J. Pollard: 1987, 'A centering approach to pronouns'. In: *Proceedings of the 25th annual meeting on Association for Computational Linguistics*. Morristown, NJ, USA, pp. 155–62, Association for Computational Linguistics.

Buschmeier, H., K. Bergmann, and S. Kopp: 2009, 'An alignment-capable microplanner for natural language generation'. In: *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*. Athens, Greece, pp. 82–89, Association for Computational Linguistics.

Byron, D., T. Mampilly, V. Sharma, and T. Xu: 2005, 'Utilizing visual attention for cross-modal coreference interpretation'. In: *Modeling and Using Context – 5th International and Interdisciplinary Conference CONTEXT 2005*. pp. 83–96.

Byron, D. K. and E. 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)*.

Byron, D. K. and L. Stoia: 2005, 'An Analysis of Proximity Markers in Collaborative Dialogs'. In: *Proceedings of the 41st annual meeting of the Chicago Linguistic Society*.

Carletta, J.: 1996, 'Assessing Agreement on Classification Tasks: The Kappa Statistic'. *Computational Linguistics* **22**(2), 249–254.

Cavicchio, F. and M. Poesio: 2009, 'Multimodal corpora annotation: Validation methods to assess coding scheme reliability'. In: M. Kipp, J.-C. Martin, P. Paggio, and D. Heylen (eds.): *Multimodal Corpora*. Springer-Verlag, pp. 109–121.

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

Dale, R.: 1989, 'Cooking up referring expressions'. In: *Proceedings of 27th Annual Meeting of the Association for Computational Linguistics*. pp. 68–75.

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

Dale, R. and J. Viethen: 2009, 'Referring Expression Generation through Attribute-Based Heuristics'. In: *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*. pp. 58–65.

Di Eugenio, B., P. W. Jordan, R. H. Thomason, and J. 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.

Diessel, H.: 2006, 'Demonstratives, joint attention, and the emergence of grammar'. *Cognitive Linguistics* **17**(4), 463–489.

Foster, M. E., E. G. Bard, M. Guhe, R. L. Hill, J. Oberlander, and A. Knoll: 2008, 'The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue'. In: *Proceedings of 3rd Human-Robot Interaction*. pp. 295–302.

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

Funakoshi, K. and S. W. T. Tokunaga: 2006, 'Group-based Generation of Referring Expressions'. In: *Proceedings of the Fourth International Natural Language Generation Conference (INLG 2006)*. pp. 73–80.

Gatt, A., A. Belz, and E. Kow: 2009, 'The TUNA-REG Challenge 2009: Overview and evaluation results'. In: *Proceedings of the 12th European Workshop on Natural Language*

Generation (ENLG 2009). pp. 174–182.

Gatt, A., I. van der Sluis, and K. van Deemter: 2007, 'Evaluating algorithms for the generation of referring expressions using a balanced corpus'. In: *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG 2007)*. pp. 49–56.

Gergle, D. and C. P. R. R. E. Kraut: 2007, 'Modeling the impact of shared visual information on collaborative reference'. In: *Proceedings of 25th Computer/Human Interaction Conference*. pp. 1543–1552.

Grishman, R. and B. Sundheim: 1996, 'Message Understanding Conference 6: A brief history'. In: *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*. pp. 466–471.

Grosz, B. J., A. K. Joshi, and S. Weinstein: 1983, 'Providing a unified account of definite noun phrases in discourse'. In: *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL 1983)*. pp. 44–50.

Grosz, B. J., A. K. Joshi, and S. Weinstein: 1995, 'Centering: A Framework for Modeling the Local Coherence of Discourse'. *Computational Linguistics* **21**(2), 203–225.

Gupta, S. and A. J. Stent: 2005, 'Automatic evaluation of referring expression generation using corpora'. In: *Proceedings of the 1st Workshop on Using Corpora in NLG*.

Halliday, M. A. K. and R. Hassan: 1976, *Cohesion in English*. Longaman.

Heeman, P. A. and G. Hirst: 1995, 'Collaborating on Referring Expressions'. *Computational Linguistics* **21**, 351–382.

Hobbs, J. R.: 1978, 'Resolving pronoun references'. *Lingua* **44**, 311–338.

Iida, R., S. Kobayashi, and T. 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*. pp. 1259–1267.

Janarthanam, S. and O. Lemon: 2009, 'Learning lexical alignment policies for generating referring expressions for spoken dialogue systems'. In: *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*. pp. 74–81, Association for Computational Linguistics.

Joachims, T.: 1998, 'Text categorization with Support Vector Machines: Learning with many relevant features'. In: *Proceedings of European Conference on Machine Learning (ECML 1998)*. pp. 137–142.

Jokinen, K.: 2010, 'Non-verbal Signals for Turn-taking and Feedback'. In: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta, pp. 2961–2967, European Language Resources Association (ELRA).

Jordan, P. W. and M. A. Walker: 2005, 'Learning Content Selection Rules for Generating Object Descriptions in Dialogue'. *Journal of Artificial Intelligence Research* **24**, 157–194.

Kameyama, M.: 1998, 'Intrasentential centering'. In: *Centering in Discourse*. Oxford University Press, pp. 89–114.

Kelleher, J., F. Costello, and J. van Genabith: 2005, 'Dynamically Structuring Updating and Interrelating Representations of Visual and Linguistic Discourse'. *Artificial Intelligence* **167**, 62–102.

Kiyokawa, S. and M. Nakazawa: 2006, 'Effects of reflective verbalization on insight problem solving'. In: *Proceedings of 5th International Conference of the Cognitive Science*. pp. 137–139.

Kranstedt, A., A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth: 2006, 'Deixis: How to Determine Demonstrated Objects using a Pointing Cone'. In: *Gesture in Human-Computer Interaction and Simulation*. Springer-Verlag, pp. 300–311.

Krippendorff, K.: 1980, *Content Analysis: An Introduction to its Methodology*. Sage Publications.

Kruijff, G.-J. M., P. Lison, T. Benjamin, H. Jacobsson, H. Zender, and I. Kruijff-Korbayova: 2010, 'Situated dialogue processing for human-robot interaction'. In: *Cognitive Systems: Final report of the CoSy project*. Springer-Verlag, pp. 311–364.

Kudo, T., K. Yamamoto, and Y. Matsumoto: 2004, 'Applying Conditional Random Fields to Japanese Morphological Analysis'. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Kuriyama, N., A. Terai, M. Yasuhara, T. Tokunaga, K. Yamagishi, and T. Kusumi: 2009, 'The role of gaze agreement in collaborative problem solving'. In: *Proceedings of the 26th Annual Conference of the Japanese Cognitive Science Society*. pp. 390–391. (in Japanese).

Mitkov, R.: 2002, *Anaphora Resolution*. Longman.

Nakatani, C. and J. Hirschberg: 1993, 'A speech-first model for repair identification and correction'. In: *Proceedings of 31th Annual Meeting of ACL*. pp. 200–207.

Noguchi, M., K. Miyoshi, T. Tokunaga, R. Iida, M. Komachi, and K. Inui: 2008, 'Multiple purpose annotation using SLAT – Segment and link-based annotation tool'. In: *Proceedings of 2nd Linguistic Annotation Workshop*. pp. 61–64.

Novak, H.-J.: 1986, 'Generating a coherent text describing a traffic scene'. In: *Proceedings of the 11th coference on Computational linguistics*. pp. 570–575.

Piwek, P. L. A.: 2007, 'Modality Choise for generation of Referring Acts'. In: *Proceedings of the Workshop on Multimodal Output Generation (MOG 2007)*. pp. 129–139.

Poesio, M., H. Cheng, R. Henschel, J. M. Hitzeman, R. Kibble, and R. J. Stevenson: 2000, 'Specifying the Parameters of Centering Theory: a Corpus-Based Evaluation Using Text from Application-Oriented Domains'. In: *ACL 2000*. pp. 400–407, Hong Kong.

Prasov, Z. and J. Y. Chai: 2008, 'What's in a gaze?: The role of eye-gaze in reference resolution in multimodal conversational interfaces'. In: *Proceedings of the 13th international conference on Intelligent user interfaces*. pp. 20–29.

Qvarfordt, P., D. Beymer, and S. Zhai: 2005, 'RealTourist–A Study of Augmenting Human-Human and Human-Computer Dialogue with Eye-Gaze Overlay'. In: M. F. Costabile and F. Paternò (eds.): *Human-Computer Interaction-INTERACT 2005*, LNCS 3585. Springer-Verlag, pp. 767–780.

Rehm, M., Y. Nakano, H.-H. Huang, A. A. Lipi, Y. Yamaoka, and F. Gruneberg: 2008, 'Creating a standardized corpus of multimodal interactions for enculturating conversational interfaces'. In: *Workshop on Enculturating Conversational Interfaces by Socio-cultural Aspects of Communication (ECI 2008)*.

Schiel, F. and H. Mögele: 2008, 'Talking and Looking: the SmartWeb Multimodal Interaction Corpus'. In: E. L. R. A. (ELRA) (ed.): *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco.

Spanger, P., M. Yasuhara, R. Iida, and T. 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)*. pp. 110 – 113.

Spanger, P., M. Yasuhara, R. Iida, and T. 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*.

Sternberg, R. J. and J. E. Davidson (eds.): 1996, *The Nature of Insight*. The MIT Press.

Stoia, L., D. M. Shockley, D. K. Byron, and E. Fosler-Lussier: 2006, 'Noun Phrase Generation for Situated Dialogs'. In: *Proceedings of the 4th International Natural Language*

Generation Conference (INLG 2006). pp. 81–88.

Stoia, L., D. M. Shockley, D. K. Byron, and E. 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)*. pp. 28–30.

Strassel, S., M. Przybocki, K. Peterson, Z. Song, and K. Maeda: 2008, ‘Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction’. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco.

Suzuki, H., K. Abe, K. Hiraki, and M. Miyazaki: 2001, ‘Cue-readiness in insight problem-solving’. In: *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*. pp. 1012 – 1017.

Tokunaga, T., C.-R. Huang, and S. Y. M. Lee: 2008, ‘Asian language resources: the state-of-the-art’. *Language Resources and Evaluation* **42**(2), 109–116.

Tokunaga, T., R. Iida, M. Yasuhara, A. Terai, D. Morris, and A. Belz: 2010, ‘Construction of bilingual multimodal corpora of referring expressions in collaborative problem solving’. In: *Proceedings of 8th Workshop on Asian Language Resources*. pp. 38–46.

van Deemter, K.: 2007, ‘TUNA: Towards a Unified Algorithm for the generation of referring expressions’. Technical report, Aberdeen University. www.csd.abdn.ac.uk/research/tuna/pubs/TUNA-final-report.pdf.

van Deemter, K., A. Gatt, and R. van Gompel and Emiel Krahmer (eds.): 2009, ‘Production of referring expressions (PRE-CogSci) 2009: Bridging the gap between computational and empirical approaches to reference’.

van der Sluis, I., P. Piwek, A. Gatt, and A. Bangerter: 2008, ‘Towards a Balanced Corpus of Multimodal Referring Expressions in Dialogue’. In: *Proceedings of the Symposium on Multimodal Output Generation (MOG 2008)*.

Vapnik, V. N.: 1998, *Statistical Learning Theory*, Adaptive and Learning Systems for Signal Processing Communications, and control. John Wiley & Sons.

Viethen, J. and R. Dale: 2008, ‘The use of spatial relations in referring expression generation’. In: *Proceesings of 5th International Natural Language Generation Conference*. pp. 59–67.

Walker, M., M. Iida, and S. Cote: 1994, ‘Japanese Discourse and the Process of Centering’. *Computational Linguistics* **20**(2), 193–232.