Understanding Referring Expressions Involving Perceptual Grouping

Kotaro Funakoshi Satoru Watanabe Takenobu Tokunaga
Department of Computer Science, Tokyo Institute Of Technology
2-12-1 Oookayama Meguro, Tokyo 152-8552, JAPAN
{koh,satoru_w,take}@cl.cs.titech.ac.jp

Naoko Kuriyama
Department of Human System Science, Tokyo Institute of Technology
2-12-1 Oookayama Meguro, Tokyo 152-8552, JAPAN
kuriyama@hum.titech.ac.jp

Abstract

This paper deals with understanding referring expressions involving perceptual grouping. The ability to use referring expressions is important for conversational agents aimed at real-world interaction. We conducted a psychological experiment to collect referring expressions involving perceptual grouping. A set of methods to identify referents based on the collected data are presented. We were able to identify 78.8% of the referents in the collected expressions.

1. Introduction

A referring expression is a linguistic product used to discriminate a specific object from the rest of the world. The ability to use referring expressions is important for conversational agents aimed at real-world interaction [1]. Referring expressions include anaphoric or exophoric expressions such as "that one" and "there", but in this paper by referring expression we mean full descriptions for identifying objects in the world such as "the red chair to the right of the table."

We have previously pointed out the importance of perceptual grouping in understanding and generating referring expressions [2]. Perceptual grouping is the human ability to recognize similar objects, or objects in close proximity to each other [6]. Most past work on generating referring expressions (*e.g.*, [4, 8]) makes use of the attributes of objects and the binary relations among them. These methods are, however, insufficient to produce natural referring expressions in some situations, as shown in Figure 1. In such a situation, algorithms that use only attributes and binary relations generate awkward expressions such as "the ball in front of a ball and in back of a ball" to designate object *c*, or even fail to generate an expression. Generating more natural descriptions, such as (*ex*1) for object *c* in Figure 1, requires recognition of similar or proximal objects, *i.e.*, perceptual grouping, and requires making use of *n*-ary relations among objects in each recognized group.

(*ex*1) the ball in the middle

In the case of the situation shown in Figure 1, both understanding and generating an expression (*ex*1) require recognizing the group consisting of objects *a*, *b*, and *c*.

Understanding referring expressions involving perceptual grouping is important in attempting to generate such expressions. This paper tackles this problem. We consider that understanding referring expressions consists of two stages: (a) semantic analysis, *i.e.*, analyzing expressions to extract semantic information, and (b) referent identification, *i.e.*, discriminating referents by using extracted information. Below we describe the experiment used to collect data and the structure of referring expressions involving perceptual grouping. We then present methods for semantic analysis and referent identification.

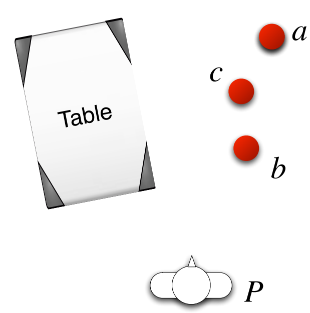


Figure 1. Example of problematic situations

2. Data collection

We conducted a psychological experiment with 42 Japanese undergraduate students to collect referring expressions in which perceptual groups are used.

2.1. Method

Subjects were presented with 2D bird's-eye images in which several objects of the same color and size were arranged. They were asked to convey a target object to a third person drawn in the same image. We used 12 images with various arrangements (see Appendix A). In each image, three to nine objects were arranged manually so that they were distributed non-uniformly. An example of the images presented to subjects is shown in Figure 2. The labels $a, \ldots, f, x$ in the figure have been assigned for the purpose of illustration and were not included in the images presented to the subjects. Each subject was asked to give a command that would enable the person in the image to pick out a target object enclosed with dotted lines. When a subject was unable to think of a suitable expression, she/he was allowed to abandon that arrangement and proceed to the next one. Referring expressions designating the target object were collected from the subjects' commands.

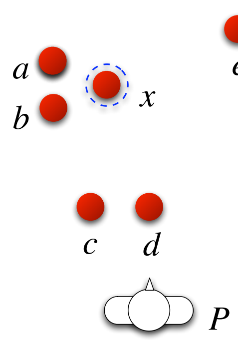


Figure 2. Visual stimulus used in the experiment

2.2. Structure of referring expressions

We presented 12 different arrangements to the 42 subjects, which should have produced 504 expressions. However, 15 judgments were abandoned and 13 expressions were apparently inadequate to identify targets. We therefore obtained 476 referring expressions. An analysis of the collected expressions showed that after starting from a group with all of the objects, subjects generally narrowed down the group to a singleton group with the target object. Therefore, a referring expression can be formalized as a sequence of groups (SOG) reflecting this narrowing-down process.

Subjects often referred to multiple groups in the course of referring to the target. In these cases, we observed two types of relations: an *intra-group relation* such as "the front two *among* the five near the desk"[1], and an *inter-group relation* such as "the two *to the right of* those five." We define an intra-group relation as a subsumption relation between two groups. Although there were two types of relations between groups, expressions using only intra-group relations made up 77% of the total.

The example below shows an observed expression describing target x in Figure 2 followed by a representation of the corresponding SOG.

(*ex*2) *hidari oku ni aru mittu no tama no uti no itiban migi no tama*
(the rightmost ball among the three balls at the back left)

SOG: $[\{a,b,c,d,e,f,x\},\{a,b,x\},\{x\}]$ [2]
where

- $\{a,b,c,d,e,f,x\}$ denotes all the objects in the image (the total set),
- $\{a,b,x\}$ denotes the three objects at the back left, and
- $\{x\}$ denotes the target.

Since narrowing down starts from the total set, the SOG representation starts with a set of all the objects and ends with a singleton group with the target. Translating the collected referring expressions into SOG representations enabled us to abstract and classify the expressions. On average, we obtained about 40 expressions for each arrangement, and classified them into 8.4 different SOG representations. The collected data are summarized in Table 1.

3. Semantic analysis

We used a simple pattern matching technique rather than full parsing to extract the necessary information from the linguistic expressions. This is because the expressions were so spontaneous and varied that building a grammar capable of handling them fully would have involved considerable expense. We acknowledge the limitations of pattern matching, and that full parsing should be considered in future. However, the method described here worked very well for the immediate goal.

The method used for semantic analysis consisted of three steps:

[1]We collected Japanese expressions, but translations are presented for non Japanese readers.

[2]We denote an SOG representation by enclosing groups with square brackets.

Table 1. Summary of collected data

Arrangement ID	1	2	3	4	5	6	7	8	9	10	11	12	Average
Number of expressions obtained	41	40	41	41	42	37	42	32	42	41	41	36	39.7
Number of different SOGs	5	6	8	8	6	12	4	15	4	11	5	17	8.4

1. Segmentation:
   segmenting referring expressions into sub-expressions, each of which describes an object group or an object in the course of referring.
2. Information extraction:
   extracting semantic information from each segmented sub-expression.
3. Relation identification:
   identifying relations between object groups indicated by sub-expressions.

All the steps were performed deterministically using character-based pattern matching. We did not use a morphological analyzer.

3.1. Segmentation

First, a referring expression is segmented into sub-expressions. This process corresponds to creating the form of an SOG: $[G_0, G_1, \ldots, G_k]$. It determines the number of groups, k, but does not identify the members of each group. G_0 is the total set and G_k is the target referent of a referring expression. This is done by finding boundaries of sub-expressions with character-based pattern matching. These boundaries are characterized by clue phrases such as "*no uchi no* (among)". Clue phrases are extracted from the collected data in section 2.

Since Japanese is a head-final language, the order of sub-expressions is in parallel with the order of groups in the SOG representation. However, the total set is often not mentioned explicitly.

For example, expression (*ex*2) in section 2 is segmented into two sub-expressions:

(*ex*3) *hidari oku ni aru mittu no tama*

(*ex*4) *no uchi no itiban migi no tama*

The boundary between these two sub-expressions is characterized by the phrase "*no uchi no* (among)". (*ex*3) corresponds to a group $\{a, b, x\}$ (*i.e.*, G_1) in Figure 2 and (*ex*4) corresponds to the target object x (*i.e.*, G_2).

3.2. Information extraction

The information types used to identify objects or groups are limited to those listed in Figure 3 due to the domain setting described in section 2. Information is extracted using regular expressions from each segmented sub-expression. We assume that one type of information appears only once in a sub-expression.

- Kinds of objects
- Cardinalities of groups
- Geometric information
- Ordinal information

Figure 3. Information types

From (*ex*3), we obtain the following information.

- The group consists of balls.
  (*tama*)
- The cardinality is three.
  (*mittu*)
- The group is at the back left.
  (*hidari oku*)

From (*ex*4), we obtain the following information.

- The group consists of balls.
  (*tama*)
- The cardinality is one.
  (No specification of quantity implies this[3].)
- The group is right-most.
  (*itiban*: first, *migi*: right)

3.3. Relation identification

As described above, each sub-expression indicates an object group (a single object is regarded as a singleton group). According to the information extracted in the previous step, each relation between two contiguous groups is identified whether it is an inter-group relation or an intra-group relation.

An intra-group relation is classified into three sub-relations:

- Directional geometric relation,
  e.g., "the *right* one *among* the three balls"
- Non-directional geometric relation,
  e.g., "the one *at the center of* the three balls"

[3] The English translation is in the singular form but the original expressions did not include information on quantity since Japanese does not have plural forms of nouns.

- Ordinal relation,
  e.g., "the *first* one *from the left*"

An inter-group relation is classified into two sub-relations:

- Directional geometric relation,
  e.g., "the ball *in front of* the three balls"
- Non-directional geometric relation,
  e.g., "the ball *near to* the three balls"

Relation identification is done using ordered rules without ambiguity. The rules were set up manually after observing the collected data.

4. Referent identification

After the semantic analysis process, referents were identified using the extracted information. In this paper, we assumed that all the participants in a situation shared an appropriate reference frame [5].

The referent identification process goes from left to right, *i.e.*, from G_1[4] to G_k (for G_i, see section 2.2 and section 3.1) with the members of groups being identified by applying an referent identification algorithm to each group. The algorithm that is applied to G_i is determined according to the relation between G_i and G_{i-1}. The identified members of G_i are used to identify the members of G_{i+1}. Like the semantic analysis process, this process is also deterministic, even though the referring expressions may be ambiguous.

Below we provide the five referent identification algorithms for the relations defined in section 3.3. Each algorithm is given as an identification function that takes as its arguments a set of objects and information to specify referents.

4.1. Intra-group relation

4.1.1. Directional geometric relation. This relation corresponds to expressions such as "at the front". The identification function is given as $Id_{intra}(C,n,\vec{d})$. C is a set of candidate objects with their coordinates expressed using the Cartesian coordinate system. C is a total set or a perceptual group specified by the sub-expression in the immediate left context, *i.e.*, G_{i-1}. n is the cardinality of the target group G_i. $\vec{d}$ is a vector representing the geometric direction, and $\vec{d}$ is given according to the reference frame. If the reference frame heads north, *i.e.*, y-axis, the vector corresponding to "right" is given as $(1,0)^T$. Our implementation has eight directions. Function $Id_{intra}(C,n,\vec{d})$ outputs a perceptual group, *i.e.*, a set of the objects with cardinality n.

[4]Because G_0 is always the total set, no processing is required.

Function $Id_{intra}(C,n,\vec{d})$ first generates all possible subsets, $Cs_1, Cs_2, \ldots, Cs_m$, with cardinality n from the candidate set C. Let a set of these subsets be Cs^n. For each subset Cs_i in Cs^n, a score $is(Cs_i)$ is given as expressed in (1). $is(Cs_i)$ represents the proximity of the members of Cs_i (a is a constant). $dist(v1,v2)$ is the Euclidean distance between two members, $v1$ and $v2$, in Cs_i.

$$is(Cs_i) = \begin{cases} \prod_{v1,v2 \in Cs_i} \exp(-a \cdot dist(v1,v2)) & |Cs_i| > 1 \\ 1 & |Cs_i| = 1 \end{cases} \tag{1}$$

$s(Cs_i,\vec{d})$, the score for Cs_i, is given by (2). $c(\vec{v},\vec{d})$ gives the component of $\vec{v}$ along $\vec{d}$, where $\vec{v}$ is the vector representing the position of a member v in Cs_i.

$$s(Cs_i,\vec{d}) = is(Cs_i) \prod_{v \in Cs_i} \frac{c(\vec{v},\vec{d})}{c_{max}^{Cs_i,\vec{d}}} \tag{2}$$

where

$$c(\vec{v},\vec{d}) = \frac{\vec{v}\cdot\vec{d}}{|\vec{d}|}$$

$$c_{max}^{Cs_i,\vec{d}} = \max_{v \in Cs_i} c(\vec{v},\vec{d})$$

Finally, $Id_{intra}(C,n,\vec{d})$ is given as in (3).

$$Id_{intra}(C,n,\vec{d}) = \operatorname*{argmax}_{Cs_i \in Cs^n} s(Cs_i,\vec{d}) \tag{3}$$

Figure 4 illustrates the nature of Id_{intra}. The two objects enclosed by the solid line are referents. Id_{intra} selects objects at the headmost side of vector $\vec{d}$ in a given set.

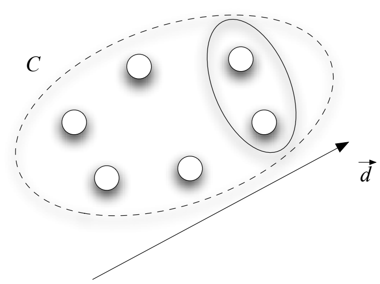


Figure 4. Illustration of Id_{intra} ($n=2$)

Suppose we try to identify objects referred to with (*ex*3) in Figure 2. As described in section 3.1, (*ex*3) corresponds to G_1. Here, C is G_0, *i.e.*, the total set $\{a,b,c,d,e,f,x\}$ (see section 2.2). $\vec{d}$ is given as a vector that represents a direction *back left* from the viewpoint of person P in Figure 2, and n is three. $Id_{intra}(C,n,\vec{d})$ gives $\{a,b,x\}$.

4.1.2. Non-directional geometric relation. This relation corresponds to expressions such as "at the center of." The identification function is given as $Ind_{intra}(C,n)$ in (4) . C and n are the same as above.

$$Ind_{intra}(C,n) = \underset{Cs_i \in Cs^n}{\operatorname{argmax}}\, is(Cs_i) \cdot js(Cs_i, c) \tag{4}$$

c in (4) is the center point of the bounding box of C. $js(Cs_i, c)$, which that represents the proximity between c and members of Cs_i, is given by (5).

$$js(Cs_i, c) = \prod_{v \in Cs_i} \exp(-a \cdot dist(v, c)) \tag{5}$$

4.1.3. Ordinal relation. The identification function is given as $Io(C, j, \vec{d})$ in (6). C is a set of candidate objects with coordinates in the Cartesian coordinate system. j is an ordinal number, and $\vec{d}$ is a vector representing the direction of geometrical ordering[5]. When ordering, we assume that a referent is a single object according to observations of the collected data. Thus, function Io outputs an object.

$$Io(C, j, \vec{d}) = \underset{v \in C}{\overset{j}{\arg}}\, c(\vec{v}, \vec{d}) \tag{6}$$

Here, $\underset{x \in S}{\overset{j}{\arg}} f(x)$ gives the elements of S that have the j-th largest value given by f.

Suppose we try to identify objects referred to with (ex4). Here, C is $\{a, b, x\}$, identified in the previous step as the referents of (ex3). $\vec{d}$ is given as a vector representing the direction *right* from the viewpoint of person P in Figure 2, and j is 1 (first). $Io(C, j, \vec{d})$ gives $\{x\}$.

4.2. Inter-group relation

4.2.1. Directional geometric relation. This relation corresponds to expressions such as "to the left of." The identification function is given as $Id_{inter}(T, r, n, \vec{d})$ in (7). T is the total set of objects in the domain, r is the reference of the inter group relation, which is identified by G_{i-1}, and n and $\vec{d}$ are the same as above.

Function $Id_{inter}(T, r, n, \vec{d})$ first generates all possible subsets, $Ts_1, Ts_2, \ldots, Ts_m$, with cardinality n from the total set T. Let the set of these subsets be Ts^n.

$$Id_{inter}(T, r, n, \vec{d}) = \underset{Ts_i \in Ts^n}{\operatorname{argmax}}\, t(Ts_i, r, \vec{d}) \tag{7}$$

where

$$t(Ts_i, r, \vec{d}) = is(Ts_i) \cdot js(Ts_i, r) \cdot ks(Ts_i, r, \vec{d})$$

[5]Although there are several types of ordinal relations in groups other than positional relations, such as size, *e.g.*, "the *biggest* one", we focused on positional order relations in this paper.

Here, $is(Ts_i)$ is given by (1). Ts_i is a subset of T, whose cardinality is n. A member v of Ts_i must satisfy $c(\vec{v} - \vec{r}, \vec{d}) > 0$[6]. $js(Ts_i, r)$ is given by (5). $ks(Ts_i, r)$ is given by (8) and represents the directional similarity between $\vec{d}$ and $\vec{v} - \vec{r}$, and $\theta^{\vec{d}, \vec{v}-\vec{r}}$ is the angle between $\vec{d}$ and $\vec{v} - \vec{r}$.

$$\begin{aligned} ks(Ts_i, r, \vec{d}) &= \prod_{v \in Ts_i} \exp(-\tan^2 \theta^{\vec{d}, \vec{v}-\vec{r}}) \qquad (8) \\ &= \prod_{v \in Ts_i} \exp\left(-\left(\frac{|(\vec{v}-\vec{r})||\vec{d}|}{(\vec{v}-\vec{r}) \cdot \vec{d}}\right)^2 + 1\right) \end{aligned}$$

Figure 5 illustrates the nature of Id_{inter}. The two objects enclosed by the solid line are referents. Id_{inter} selects objects that are near to r and in the direction of $\vec{d}$.

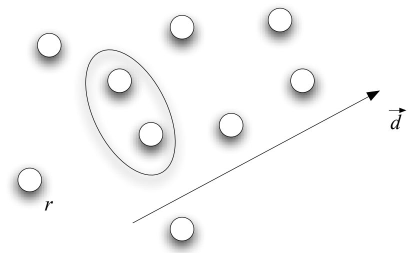


Figure 5. Illustration of Id_{inter} ($n = 2$)

is, *js* and *ks* were designed based on the potential functions defined earlier [7]. These potential functions represent spatial plausibilities. The value of parameter a was determined to 0.001 experimentally.

4.2.2. Non-directional geometric relation. This relation corresponds to expressions such as "near to." The identification function is given as $Ind_{inter}(T, r, n)$ in (9) . T, r and n are the same as above.

$$Ind_{inter}(T, r, n) = \underset{Ts_i \in Ts^n}{\operatorname{argmax}}\, is(Ts_i) \cdot js(Ts_i, r) \tag{9}$$

$is(Ts_i)$ and $js(Ts_i, r)$ are given by (1) and (5) respectively.

(9) and (4) are almost the same but there is a difference between the arguments.

5. Experimental results and discussion

5.1. Results

We implemented the methods describe in section 3 and section 4 in Perl and applied them to the expressions collected in the experiment.

[6]If the relation is "next to", this constraint is not required, because "next to" is bidirectional.

Table 2 shows the results for all relations (intra-group and inter-group relations). The first row, Total, shows the results for all collected expressions. The second row, Applicable, shows the results only for expressions that the proposed methods should be able to handle. The second column of each table, Pop., shows the number of expressions. The third column, Analysis, shows the success rate of the semantic analysis. The fourth column, Ident., shows the success rate of the referent identification. The number of Ident. under Total in Table 2 shows the overall performance of the method.

In the 425 'Applicable' expressions, there were 175 boundaries to be segmented. Our method found 168 boundaries, and 164 of these were correct. Thus, the precision rate for segmentation was 97.6% (164/168) and the recall rate for segmentation was 93.7% (164/175).

Referent identification succeeded even with 15 expressions out of 51 not classified as Applicable. This gave the 3.5% gain in Ident. under Total in Table 2. These successes are creatures of chance from the theoretical viewpoint in the sense that our methods did not fully utilize information in those expressions. However, they can be regarded as demonstrating the robustness of our methods.

Table 2. Experimental results: all expressions

Expressions	Pop.	Analysis	Ident.
Total	476	83.0%	78.8%
Applicable	425	92.9%	84.7%

Table 3 shows the results for expressions including intra-group relations only, and Table 4 shows the results for expressions including both intra- and inter- group relations.

Table 3. Experimental results: expressions without inter-group relations

Expressions	Pop.	Analysis	Ident.
Total	368	85.0%	84.8%
Applicable	332	94.3%	90.1%

Table 4. Experimental results: expressions with inter-group relations

Expressions	Pop.	Analysis	Ident.
Total	108	75.9%	58.3%
Applicable	93	88.1%	65.6%

We asked 42 human subjects to identify referents in the 117 expressions in the 476 expressions collected in section 2. They were completely different from those used in section 2. Each expression was assigned to about 14 subjects. Table 5 compares the success rate of humans subjects and our method in identifying referents in the 117 expressions. As these results show, the performance of our method was similar to that of the human subjects. The third row of the table shows that the machine performance was superior to human performance. However, because there were only a small number of cases, the difference was not significant.

Table 5. Human vs. machine performance

Expressions	Pop.	Human Ident.	Machine Ident.
Total	117	87.3%	80.3%
w/o Inter rel.	84	94.4%	83.3%
w/ Inter rel.	33	69.1%	72.7%

5.2. Error analysis

5.2.1. Errors in semantic analysis. Most errors in semantic analysis are due to non-linearity of referring. As described in section 2.2, the SOG representation presupposes linearity of referring. Thirty eight expressions in the collected data include non-linear referring such as "the ball in front of the table and in back of the chair".

Among these expressions, 22 include exclusive expressions; for example, "the four balls except for the front-most ball and the back-most ball." Ten expressions are composed of multiple sentences. Such expressions also exceed the capability of the SOG representation.

5.2.2. Errors in referent identification. Some expressions, such as "the center ball among the bunch of balls to the right", indicate a group of objects without mentioning their cardinality. It is hard to specify a set without using cardinality information (*e.g.*, "a bunch of balls to the right"). However, humans seem able to identify the targets of these expressions by using perceptual grouping. Previously proposed methods of simulating human perceptual grouping, *e.g.*, [6], would help in such situations.

Thirteen expressions require recognition of salient geometric formations, *e.g.*, line. Of these 13 expressions, 8 refer to lines and the rest refers to other shapes. This cannot be handled even by the perceptual grouping method mentioned above [6]. We therefore need a mechanism capable of recognizing such specific formations.

6. Conclusion

We studied referring expressions involving perceptual grouping. This was performed in two stages: semantic analysis and referent identification. Referring expressions that were collected experimentally were analyzed using a simple pattern-matching technique. This method succeeded in analyzing 83.0% of the collected expressions. Using the information extracted from the expressions by the proposed semantic analysis method, 78.8% of the targets were correctly identified.

Although the method was applied to very limited information in this paper, we believe that it could easily extended to other types of information such as the color and size of objects.

We also believe that these results demonstrate sufficient understanding of single referring expressions and that further efforts will produce little improvement. In future work, it would be interesting to combine the proposed methods with interactive or collaborative systems of object identification. In interactive object identification, the user and system incrementally identify objects through dialogue using multiple utterances (see the sample dialogue (*ex*5)). This type of functionality would considerably increase the usability of conversational agents. The previously proposed computational model of collaborating on referring expressions [3] would be a good basis.

(*ex*5) An example of interactive object identification

U: Do you see a round table?

S: Yes.

U: There are balls in front of it.

S: Three red balls, right?

U: Right.

U: Take the left one.

S: OK.

Acknowledgment

This work is partially supported by a Grant-in-Aid for Creative Scientific Research 13NP0301, the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] Donna K. Byron. Understanding referring expressions in situated language: Some challenges for real-world agents. In *the First International Workshop on Language Understanding and Agents for the Real World*, 2003.

[2] Kotaro Funakoshi, Satoru Watanabe, Naoko Kuriyama, and Takenobu Tokunaga. Generation of relative referring expressions based on perceptual grouping. In *the 20th International Conference on Computational Linguistics*, 2004.

[3] Peter Heeman and Graem Hirst. Collaborating referring expressions. *Computational Linguistics*, 21(3):351–382, 1995.

[4] Emiel Krahmer, Sebastiaan van Erk, and André Verleg. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72, 2003.

[5] Stephen C. Levinson, editor. *Space in Language and Cognition*. Cambridge University Press, 2003.

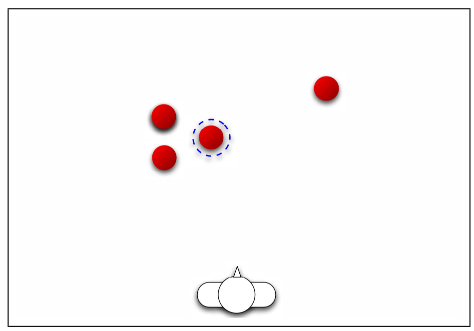
[6] K. R Thórisson. Simulated perceptual grouping: An application to human-computer interaction. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pages 876–881, 1994.

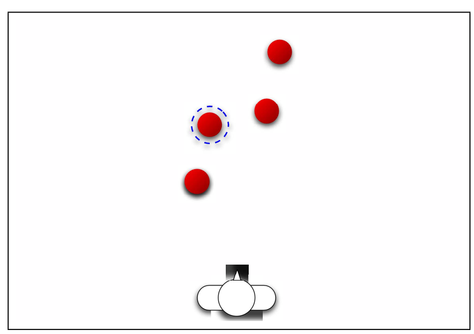
[7] Takenobu Tokunaga, Tomofumi Koyama, and Suguru Saito. Meaning of Japanese spatial nouns. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimentions of Prepositions and their Use in Computational Linguistics: Formalisms and Applications*, pages 93–100, 2005.

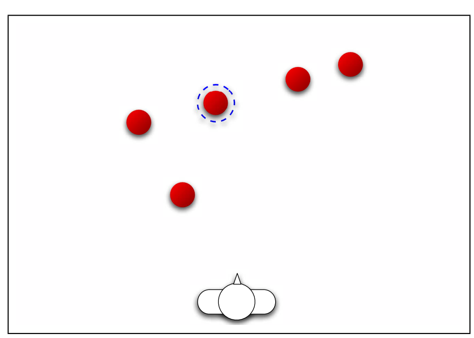
[8] Kees van Deemter. Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52, 2002.

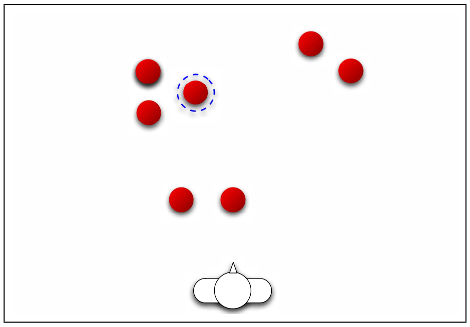
A. Arrangements

The twelve arrangements used in the data collection experiment are shown below.

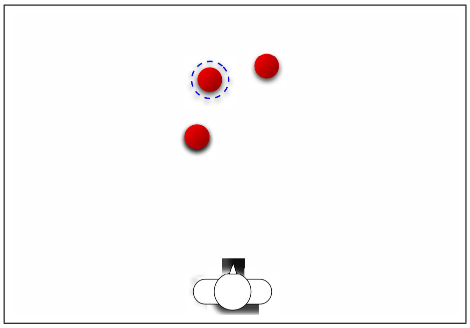
1.


2.


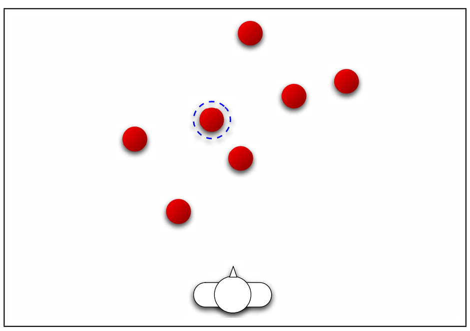
3.




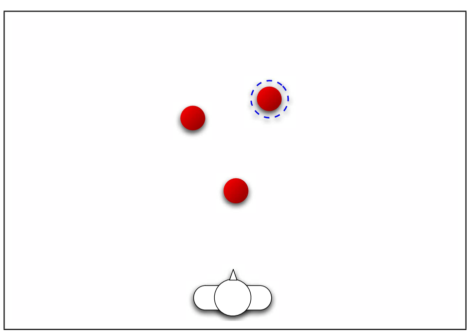
4.



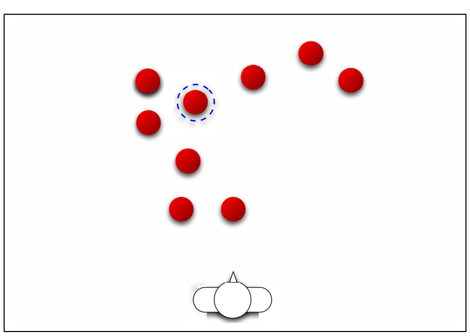
5.



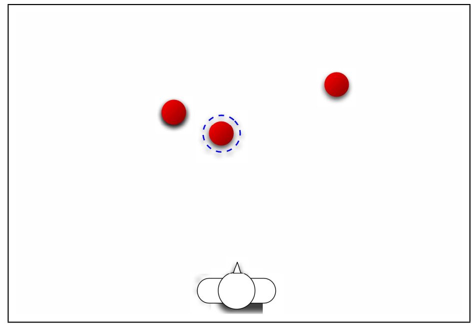
6.



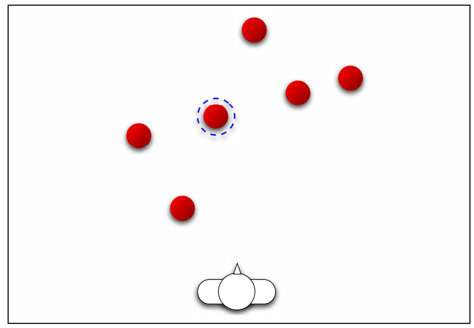
7.



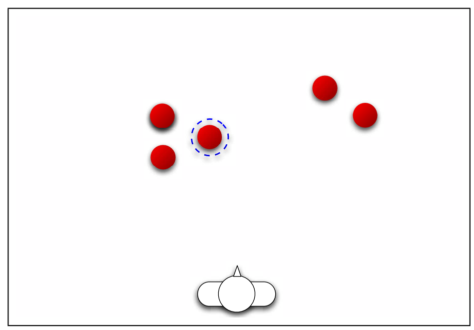
8.



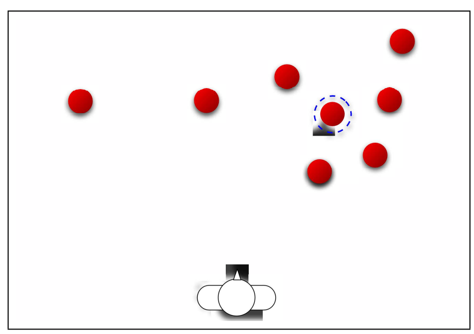
9.



10.



11.



12.