Vol.2012-NL-207 No.5
2012/7/26

Comparative study of generating referring expressions in situated dialogues

Bibyna Feraena[1] Iida Ryu[1] Tokunaga Takenobu[1]

Abstract: This paper focuses on generating referring expressions (REs) using the REX corpora, which consist of the set of dialogues where two participants collaboratively solve Tangram puzzles. In this study, we show our manual analysis of the different use of the REs in English and Japanese corpora. In addition, we report the results of our empirical evaluation replicating the usage of demonstrative pronouns in the corpora by machine learning techniques.

1. Introduction

The effective use of referring expressions, as a linguistic means to refer to one and more specific objects in a situation, is fundamental to a smooth collaboration between humans and computers. While the task of referring expression generation might be a trivial task for humans, it poses significant challenge to create a system that is able to replicate this task. Initial research on the generation of referring expressions focused on generating isolated expression in static domains (e.g. images) to distinguish a certain object from the distractors using its attributes [3], [4]. However, such environments do not necessarily reflect actual human reference behaviour, where the collaborative aspect plays a central role and the situation is constantly changing. In fact, this dynamic behaviour has been pursued since the early study on [2], [6].

In addition, there has been a shifting in recent research interest towards studying the multimodal phenomenon where extra-linguistic information takes part to generation of referring expressions aside from linguistic information. There have been attempts to develop algorithms combining these two kinds of information into a computational model to generate referring expressions. Research in referring expressions in a situated collaborative dialogue, where the participants can impact their environment, is important in this research trend.

Realising the importance of collaboration between extra-linguistic information and linguistic information in order to generate appropriate referring expressions, Spanger *et al.* constructed a multimodal Japanese corpus (REX-J) to build a computational model to replicate the generation of referring expressions by human, and evaluated the system by using the collected corpus [8]. The English counterpart of this corpus has also been constructed and analysed [9].

The purpose of this paper is to replicate the experiment by Spanger *et al.* using the English REX corpus (T2010-03) and to analyse the difference between these evaluation results. The analysis results show that while in some respect there are similarities, there are also some interesting differences between the two corpora.

This paper is organized as follows. Section 2 provides an overview of previous work on generation of referring expressions. Section 3 explains the corpora used in the experiment. Section 4 discusses the experiment itself. Section 5 presents the conclusion of this experiment.

2. Related Work

Generation of referring expressions (GRE) is the part of research area of Natural Language Generation (NLG). NLG might be considered as the opposite of Natural Language Understanding (NLU), but while NLU deals with a problem of hypothesis management, NLG deals with a problem of choice. As noted in Section 1, a referring expression is a fundamental device in communication.

GRE can be distinguished based on whether they are used in static or dynamic environment. In static environment, referring expressions are used in an invariant situation, such as image, in order to distinguish the target object from distractors in the domain. There are two different approaches to the design of generation algorithm: a rule-based approach and an empirical approach. The seminal work in this field, the Incremental Algorithm [4], is based on the former approach, which used a set of attributes incrementally in content determination of the target object. However, as stated in Section 1, static environment does not necessarily represent the real environment.

Dynamic environment ranges from simple text dialogues to more realistic setting such as *situated* dialogues, where the participants share space and can act upon objects in

[1] Department of Computer Science,
Tokyo Institute of Technology

that space. Generation of referring expressions in a situated dialogue needs extra-linguistic information as well as pure linguistic information. It has also been noted that in a collaborative task, the participant's actions on object heavily influenced their referential behaviour [5]. Spanger *et al.* proposed a GRE model in a situated dialogue that focused on integration of linguistic and extra-linguistic information, particularly information of physical action[8].

In order to able to generate referring expressions in a realistic domain such as situated dialogue, the construction of multimodal corpora is a critical task. An initial study was the construction of Map Task Corpus [1]. The REX corpora were constructed from dialogues between pairs that were given a goal to solve the Tangram puzzles collaboratively. The set-up was chosen so that the two participants share the same task space, but has different role and action that they can perform.

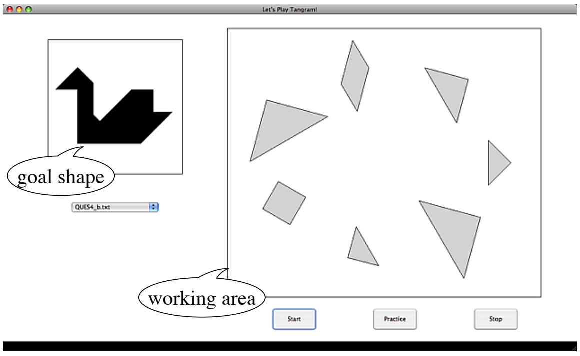

Fig. 1 Tangram puzzle

3. Corpora

The corpora used in this study were constructed from collecting dialogues from participants that were given a task to solve the Tangram Puzzles within a 15 minutes time limit [9]. The Tangram puzzle is a geometrical puzzle that originated in ancient China. The goal of a Tangram puzzle is to construct a given goal shape by arranging seven pieces that include two large triangle, a medium-sized triangle, two small triangles, a parallelogram and a square. An example of simulator that was used to construct the corpora is shown in **Fig. 1**.

Participants are separated into pairs, and each individual in a pair is given one of two different roles. One acts as a solver (SV), while the other acts as operator (OP). Only the solver has the goal shape in the screen, but only the operator has mouse for manipulating the pieces. This asymmetrical setting requires collaboration between the two to achieve the goal in hand: the solver has to think and instruct the operator, while the operator has to manipulate the pieces based on solver's instruction.

This study used corpus named T2008-08 for Japanese and T2010-03 for English from the REX corpora [9]. Each corpus contains 24 dialogues by 6 pairs of native speakers. The recorded speech was transcribed and the referring expressions were annotated using ELAN[*1], a multimodal annotation tool. ELAN was also used to merge the annotation with extra-linguistic data recorded by the Tangram simulator. Extra-linguistic information consists of the action on a piece, the movement of mouse cursor, and the position of each piece in the working area. Further detailed description of the corpora can be found in [9].

Table 1 Number of utterances and referring expressions

corpus	#utterances		#referring exp.	
	OP	SV	OP	SV
Japanese (T2008-08)				
Total	1,892	2,571	200	1,214
Ave.	78.8	107.1	8.3	50.6
SD	51.0	40.6	10.4	19.3
English (T2010-03)				
Total	2,049	4,848	310	2,396
Ave.	85.4	202.0	12.9	99.8
SD	64.0	70.1	10.2	42.5

Intuitively, the solver would use more referring expressions because they have to give direction to the operator, while the operator usually use referring expressions to confirm the instruction or to ask for more information. This is confirmed for dialogues in both English and Japanese corpora, as can be seen in **Table 1**. The row "Ave." shows the average numbers over 24 dialogues and the row "SD" shows their standard deviations.

Table 2 Attributes of referring expressions

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

There are different types of attributes of referring expressions that are used in the corpus annotation. **Table 2** shows the description of each attribute, and **Table 3** shows their distribution in the Japanese and English corpora. In the Japanese corpus, the demonstrative pronoun is the second most frequent type of referring expressions, following the expressions utilising intrinsic attributes (siz, typ, dir) of pieces. In contrast, the demonstrative pronoun is the most frequent type of referring expressions used in the English corpus.

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

Table 3 Number of referring expression attributes

corpus	dpr	dad	dmn	siz	col	typ	dir	prj	tpl	ovl	act	cmp	sim	num	rpr	err	nest	meta	nul
Japanese (T2008-08)																			
Total	668	176	39	285	0	647	7	141	10	2	94	29	7	35	2	2	30	6	10
[%]	23.1	6.1	1.3	9.9	0	22.4	0.2	4.9	0.3	0.1	3.3	1.0	0.2	1.2	0.1	0.1	1.0	0.2	0.3
English (T2010-03)																			
Total	1,835	374	0	422	35	725	2	132	40	7	48	144	0	79	7	5	24	22	10
[%]	46.9	9.6	0	10.8	0.9	18.5	0.1	3.4	1.0	0.2	1.2	3.7	0	2.0	0.2	0.1	0.6	0.6	0.3

4. Experiments

Following the Spanger *et al.* [8], we replicate their experiment of generating demonstrative pronouns in appropriate contexts by using both English and Japanese corpora. The task of generation of demonstrative pronouns is addressed in terms of deciding whether to use demonstrative pronoun or not in a certain situation. The model will take a situation and a target object as input, and will output whether to use a demonstrative pronoun or not. The aim of this experiment was to replicate the human's usage of demonstrative pronouns in the corpus, assuming that an expression that has been actually used by human in a certain situation can be considered as a correct answer. This experiment employed a machine learning approach utilising both linguistic and extra-linguistic information. This information is employed as features for a Support Vector Machine (SVM) [10]. The features represents an input situations for generating a referring expression. Given a set of training examples which are classified as either positive or negative class in advance, the SVM seeks a separating hyperplane by maximising the margin between the two classes. In this case, the SVM has to decide whether to use demonstrative pronoun (positive) or not (negative). In this experiment, we utilised the SVM-light software [7]. To simplify the experiment, only the referring expressions referring to a single target piece were considered. Because the size of the data is small, we conducted the 10-fold cross validation.

Table 4 Features representing a situation for a referring expression

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

4.1 Features

As input for the SVM, we need to define the feature-vectors representing a situation when the target is mentioned using a referring expression. We follow Spanger *et al.* [8] to use the 12 features shown in **Table 4**. There are three categories of the features: discourse history features (D1–D5), action history features (A1–A5), and the current operation features (O1 and O2).

The dialogue features model the linguistic information while the action and operation features model the extra-linguistic information from the collaboration, which might have an impact on the usage of demonstrative pronouns. These features then was split out again according to their respective values, which were resulted in 27 features like shown in **Table 7**. The aim here is to automatically decide whether or not to use demonstrative pronouns to refer to the target piece in a certain situation represented by these features.

Table 5 Result of classification (all instances)

	Japanese			English		
Targeting	R	P	F	R	P	F
DP	0.785	0.780	0.783	0.863	0.849	0.856
non-DP	0.790	0.795	0.792	0.633	0.660	0.646

4.2 Results and Discussion

Given a situation and a target object, the SVM classifier classifies the case into two classes: "demonstrative pronoun (DP)" and "other (non-DP)". **Table 5** shows recall (R), precision (P) and F-measure (F) using all features. In total, there are 1,224 instances (596 DPs and 628 non-DPs) in the Japanese corpus and 2,362 instances (1,664 DPs and 698 non-DPs) in the English corpus. Clearly, the English corpus is skewed to the DP instances, so it would naturally results in a higher performance when targeting DPs, i.e. considering DPs as positive instances. Another experiment was conducted using a balanced corpus where instances of both classes are randomly sampled so that the number of instances in each of them is equal to 596, which is the number of instances with DPs in the Japanese corpus.

Table 6 Result of classification (balanced instances)

	Japanese			English		
Features	R	P	F	R	P	F
All	0.789	0.785	0.786	0.795	0.752	0.772
w/o D1–D5	0.786	0.785	0.784	0.768	0.733	0.749
w/o A1–A5	0.786	0.785	0.784	0.768	0.733	0.749
w/o O1, O2	0.719	0.689	0.698	0.759	0.700	0.727

Table 6 shows the overall result of the classification. We conducted the feature ablation by excluding features of each feature categories, i.e. discourse history, action history and current operation. The row "All" shows the result of classification when all features are used. The succeeding rows show the results of classification when a feature category

Table 7 Learnt weight of features using (all instances)

	Japanese		English	
Rank	Feature	Weight	Feature	Weight
1.	O2=target	0.8655	O2=target	0.4931
2.	D1≤10	0.1408	A1≤10	0.4467
3.	A5=yes	0.0664	D1≤10	0.4140
4.	D2=pron	0.0595	A5=yes	0.2754
5.	O1=target	0.0568	D5=yes	0.2426
6.	D5=yes	0.0563	D2=pron	0.1172
7.	A1≤10	0.0510	D2=other	0.1169
8.	D4≤20	0.0438	A2=rotate	0.0926
9.	A4≤20	0.0349	A2=move	0.0924
10.	A4>20	0.0337	A2=flip	0.0915
11.	A2=flip	0.0260	D4≤20	0.0366
12.	D2=other	0.0164	D4≤10	0.0366
13.	A2=rotate	0.0084	D4>20	0.0363
14.	D4>20	0.0049	A4>20	0.0203
15.	D4≤20	0.0034	A4≤10	0.0200
16.	D3	0.0021	A4≤20	0.0200
17.	A3	0.0003	O1=target	0.1480
18.	A4≤10	-0.0046	D3	0.0005
19.	A1>20	-0.0051	A3	0.0005
20.	D1≤20	-0.0210	O1=no	-0.0138
21.	A2=move	-0.0331	D1≤20	-0.0627
22.	A5=no	-0.0382	A1≤20	-0.0627
23.	D1>20	-0.0433	A1>20	-0.1073
24.	A1≤20	-0.0444	D1>20	-0.1165
25.	D5=no	-0.0553	A5=no	-0.1564
26.	O1=no	-0.5580	D5=no	-0.2416
27.	O2=no	-0.8645	O2=no	-0.4921

Table 8 Learnt weight of features (balanced instances)

	Japanese		English	
Rank	Feature	Weight	Feature	Weight
1.	O2=target	0.8344	O2=target	0.3225
2.	D1≤10	0.1793	D1≤10	0.2620
3.	A5=yes	0.1149	A5=yes	0.2567
4.	O1=target	0.0596	D2=pron	0.2444
5.	D2=pron	0.0592	A1≤10	0.1969
6.	D5=yes	0.0584	D5=yes	0.1487
7.	A1≤10	0.0456	O1=target	0.0785
8.	D4≤10	0.0420	A2=rotate	0.0722
9.	A4≤20	0.0389	D4≤10	0.0631
10.	A4>20	0.0378	A4≤20	0.0382
11.	D2=other	0.0372	A4>20	0.0382
12.	A2=flip	0.0286	A2=flip	0.0184
13.	D4≤20	0.0198	D4≤20	0.0182
14.	D4>20	0.0198	A3	0.0005
15.	A4≤10	0.0165	D3	0.0005
16.	A1>20	0.0030	D4>20	-0.0320
17.	D3	0.0011	D1≤20	-0.0342
18.	A2=rotate	0.0005	A4≤10	-0.0569
19.	A3	0.0003	A2=move	-0.0733
20.	A2=move	-0.0223	O1=no	-0.0775
21.	D1≤20	-0.0377	D1>20	-0.0788
22.	A1≤20	-0.0418	A1≤20	-0.0893
23.	D1>20	-0.0446	A1>20	-0.0902
24.	A5=no	0.0573	D2=other	-0.0958
25.	D5=no	-0.0574	D5=no	-0.1477
26.	O1=no	-0.0586	A5=no	-0.2198
27.	O2=no	-0.8334	O2=no	-0.3125

Table 9 Results of feature combinations (Japanese)

ID	feature	R	P	F
1.	O2=target	0.789	0.785	0.786
2.	D1≤10	0.789	0.785	0.786
3.	A5=yes	0.789	0.785	0.786
4.	O1=target	0.789	0.785	0.786
5.	D2=pron	0.789	0.785	0.786
6.	D5=yes	0.789	0.785	0.786
7.	A1≤10	0.789	0.785	0.786
8.	D4≤10	0.789	0.785	0.786
9.	A4≤20	0.789	0.785	0.786
10.	A4>20	0.789	0.785	0.786
11.	D2=other	0.789	0.785	0.786
12.	A2=flip	0.789	0.785	0.786
13.	D4≤20	0.789	0.785	0.786
14.	D4>20	0.789	0.785	0.786
15.	A4≤10	0.789	0.785	0.786
16.	A1>20	0.789	0.785	0.786
17.	D3	0.789	0.785	0.786
18.	A2=rotate	0.789	0.785	0.786
19.	A3	0.783	0.805	0.793
20.	A2=move	0.783	0.805	0.793
21.	D1≤20	0.783	0.804	0.792
22.	A1≤20	0.789	0.804	0.795
23.	D1>20	0.783	0.799	0.789
24.	A5=no	0.786	0.809	0.796
25.	D5=no	0.787	0.804	0.795
26.	O1=no	0.789	0.805	0.796
27.	O2=no	0.789	0.785	0.786

Table 10 Results of feature combinations (English)

ID	feature	R	P	F
1.	O2=target	0.768	0.733	0.749
2.	D1≤10	0.768	0.732	0.754
3.	A5=yes	0.770	0.729	0.749
4.	D2=pron	0.770	0.730	0.750
5.	A1≤10	0.768	0.733	0.749
6.	D5=yes	0.768	0.733	0.749
7.	O1=target	0.768	0.733	0.749
8.	A2=rotate	0.768	0.733	0.749
9.	D4≤10	0.768	0.733	0.749
10.	A4≤20	0.768	0.733	0.749
11.	A4>20	0.768	0.733	0.749
12.	A2=flip	0.768	0.733	0.749
13.	D4≤20	0.768	0.733	0.749
14.	A3	0.795	0.748	0.770
15.	D3	0.790	0.740	0.763
16.	D4>20	0.789	0.738	0.762
17.	D1≤20	0.791	0.735	0.761
18.	A4≤10	0.793	0.737	0.763
19.	A2=move	0.793	0.737	0.763
20.	O1=no	0.794	0.728	0.759
21.	D1>20	0.795	0.738	0.765
22.	A1≤20	0.795	0.738	0.765
23.	A1>20	0.795	0.734	0.762
24.	D2=other	0.786	0.715	0.749
25.	D5=no	0.772	0.733	0.752
26.	A5=no	0.765	0.705	0.743
27.	O2=no	0.795	0.752	0.772

is removed. Interestingly, in both corpora, the F-measure without action history features and dialogue history features are the same. The F-measure when not using current operation features is the lowest, which reflects the fact that information on ongoing action has a strong impact on the use of DPs.

Table 7 shows the ranked list of the learnt weight of each feature when using all features and all instances for both corpora. The weight of a feature reflects its importance in determining the classification result. **Table 8** shows the rank of features when using the balanced instances. There are some differences in the result between these two tables, but the tendencies are the same.

In both tables, feature “O2=target” topped the rank both in Japanese and English corpora. In both corpora, the operation features are relatively high ranked, which means that the operation feature plays an important role on the decision whether or not to use DPs. Out of 1,664 DP instances in English corpus, only 300 instances does not have a positive value for the feature “O2=target”. In other words, most mention of a piece using a DP occurs when the mouse cursor was on the target piece. By looking into the dialogues, we also found that in many occurrences of DPs when the target is not under the mouse happened when the target piece is either recently mentioned or operated. This supports the high weight of the action and discourse features like "D1≤10", "A5=yes" and "A1≤10" in both Table 7, Table 8. This further strengthens the point that linguistic and extra-linguistic information have to be integrated for the generation of appropriate referring expressions.

In the Japanese corpus, the gap of weight between the first and the second ("D1≤10" in both cases) is quite big. This means that in Japanese corpus, the use of DP is determined by almost only the feature "O2=target", i.e. the mouse cursor is on the target. Interestingly, this gap is much smaller in the English corpus. The differences of weights are quite small among top four features. This means that unlike its Japanese counterpart, some instances in the English corpus need a complex combination of features in order to decide whether or not to use DP. We have no reasonable explanation for this difference at this moment.

Based on the feature weights shown in Table 8, we investigated the impact of each feature by evaluating the performance of feature combinations, which was generated by adding one feature at a time in descending order. Feature combination ID M includes the first ranked feature through the M-th ranked feature. **Table 9** and **Table 10** show the F-measure, precision and recall over feature combinations 1-27 in both corpora.

In the Japanese corpus, the F-measure stays constant from the feature combination 1 to 18. This reflects the result in Table 8, where the weight of "O2=target" has a particularly large gap from the other features' weight, making it the most influential feature to indicate the use of DPs. Interestingly, using all features in this corpus is not the best combination in terms of F-measure.

In contrast, in the English corpus, the F-measure grows from the feature combinations 1 and 2, but remains relatively the same until the feature combination 13, where it grows when added feature "A3". It stayed relatively the same again and dropped when added feature "A5=no", reaching the highest score when using all features.

We investigated the errors produced both in Japanese and English corpora by investigating the results of the feature combination ID 2 for both corpora. Adding more features after this feature combination did not give much improvment of the F-measure. There are two types of errors: false positives (FP: when humans do not use DPs but the classifier predictsa DPs), and false negatives (FN: when humans uses DPs but the classifier predicts non-DP). FPs negatively impact precision and FNs negatively impact recall.

There are 132 cases of FPs and 160 cases of FNs in balanced data set of the English corpus. The FP cases occurred when either one of the used feature (O2=target, D1≤10) has a positive value. FN cases happened when one of these features has a negative values. We observed that there are several cases where the participants use DPs even though the mouse is not on the target. In many cases, the solvers mentioned other pieces which has some relation with the pieces under current operation, e.g. mentioning piece 6 as the operator was moving piece 3 towards piece 6. There is also many cases where the mouse cursor is not actually on any piece, or located near the mentioned pieces. Even though the feature "D1lew10" is positive, there are cases that are not able to be captured in our model, where the participants mentioned two pieces together and later refered again to one of them.

In the Japanese corpus, there are 132 cases of FPs and 225 cases of FNs. Our main observation is that all cases of FPs happen when the feature "O2=target" has a positive value, i.e. the target is under mouse cursor at the time of the referring expression. All of the FNs cases is the complete opposite of this, they happened when this feature has a negative value. This further shows how dominant the operation feature is in our model, especially for the Japanese corpus.

5. Concluding remark

This paper compared the performance of the GRE algorithm across the Japanese and English corpora of situated dialogues for collaborative problem solving. Following Spanger *et al.* [8], we particularly focused on generating demonstrative pronouns in a multimodal problem solving setting. The corpora used in this study were constructed by collecting dialogue in each language, where participants are asked to collaboratively solve Tangram puzzles, with a 15 minutes time limit.

We made SVM classifiers that decide whether a demonstrative pronoun is appropriate to refer to a given target puzzle piece in a given situation. The situation was represented by both linguistic and extra-linguistic information including actions on a piece, the position of mouse cursor. The feature ablation revealed that the feature indicating the mouse was on the target piece at the use of the referring expression ("O2=target") had a dominant impact on the use of demonstrative pronouns in both Japanese and English corpora. However, a more precise analysis showed that the gap of feature weights between this top feature and the succeeding ones was smaller in the English corpora than in the Japanese one. That is, "O2=target" is a almost decisive feature for the use of demonstrative pronouns in Japanese. We have no reasonable explanation for this difference at this moment.

We also performed an experiment using various features combinations, i.e. using the features accumulatively one by one based on their rank on the learnt weight. As a natural consequence of the above discussion, using only the feature "O2=target" resulted in the best result in the Japanese corpus, while in the English corpus, using this feature together with "D1≤10" resulted in the best result.

References

[1] Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weiniert. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366, 1991.

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

[3] Robert Dale. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 68–75, 1989.

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

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

[6] Peter Heeman. Combining reinformation learning with information-state update rules. In *Proceedings of Human Language Technologies 2007/ The Conference of the North American Chapter of the Association for Computational Linguistics (ACL 2007)*, pages 268–275, 2007.

[7] Thorsten Joachims. Making large-scale SVM learning practical. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 169–184. MIT-Press, 1999.

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

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

[10] Vladimir N. Vapnik. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing Communications, and control. John Wiley & Sons, 1998.