A Cross-Linguistic Study on the Production of Multimodal Referring Expressions in Dialogue

Ielka van der Sluis
Communication and Information Studies
University of Groningen, the Netherlands
i.f.van.der.sluis@rug.nl

Saturnino Luz
Department of Computer Science
Trinity College Dublin, Ireland
luzs@scss.tcd.ie

Abstract

This paper presents a cross-linguistic data elicitation study on fully realised referring expressions (REs) in a dialogue context. A web-based experiment was set up in which participants were asked to choose REs to be uttered by one of two agents for identifying five targets in a scripted dialogue. Participants were told that the agent would point at the referents while uttering their chosen linguistic descriptions. The study was conducted in English, Japanese, Portuguese and Dutch and yielded a total of 1190 referring expressions. Our hypotheses concern sets of objects that need to be considered for identification depending on the effect of the pointing gesture. Results show interesting and significant differences between the language groups.

1 Introduction

Generation of referring expressions (GRE) has been a central task in Natural Language Generation for many years, and numerous algorithms which automatically produce referring expressions (REs) have been developed (Gardent, 2002; Krahmer et al., 2003; Jordan and Walker, 2005; Van Deemter, 2006). Existing GRE algorithms generally assume that both speaker and addressee have access to the same information. In most cases this information is represented by a knowledge base that contains the objects and their properties present in the domain of conversation in terms of attribute-value pairs. A typical algorithm (Dale and Reiter, 1995) takes as input an object or a set of objects (Van Deemter, 2002), the target referent of the description, and a set of distractors from which the target needs to be distinguished. The task of a GRE algorithm is to determine which set of properties is required to single out the target from the distractors.

Much of the work on GRE focusses on the use of REs in the English language. However, in recent years, other languages have attracted increased interest (Funakoshi et al., 2006; Pareira and Paraboni, 2008; Spanger et al., 2009; Theune et al., 2010). In this paper we present a cross-linguistic study on human production of REs in English, Japanese, Dutch and Brazilian Portuguese. The study originated from a project in which the perception of multimodal REs was studied in a virtual world in a Japanese and an English-speaking setting (Van der Sluis and Luz, 2011; Van der Sluis et al., to appear). In the present paper, the materials from a production study initially conducted for Japanese to validate our Japanese translation of a dialogue written in English, have been translated and further adapted to Dutch and Portuguese. We draw on the results of this study to analyse how well different languages match a typical GRE algorithm that uses a list of preferred properties, such as the algorithm proposed by Dale and Reiter (1995).

The REs considered in this study are part of a scripted dialogue between two agents in a furniture sales setting. The study focusses on 'first-mention' REs that identify objects that have not been talked about earlier in the discourse. In the dialogue the furniture seller agent refers to objects in the domain by uttering each scripted RE combined with a pointing gesture directed to the target. Since human com-

Proceedings of the 13th European Workshop on Natural Language Generation (ENLG), pages 53–62,
Nancy, France, September 2011.

munication includes gestures as well as language various algorithms for the generation of such multimodal REs have been proposed (André and Rist, 1996; Kranstedt et al., 2006; Van der Sluis and Krahmer, 2007). Interestingly, we know from other studies (Piwek, 2009; Van der Sluis and Krahmer, 2007) that the use of pointing gestures can have a particular influence on the REs in that they reduce the distractor set such that often less properties are needed to uniquely distinguish the target. In this paper we test two hypotheses about the composition of the distractor set.

The paper is structured as follows: Section 2 describes the materials and setting of the study, Section 3 presents our hypotheses and our evaluation method, Section 4 details the results, Section 5 discusses the findings and Section 6 concludes the paper.

2 Production Study

2.1 Setting: Dialogue and REs

A dialogue script was written by hand for two agents in a furniture store. Figure 1 presents a schematic layout of the furniture shop marking the positions of the agents and the furniture items. The shop contains 26 objects of which 14 were used as target referents, the others were used as distractors. The dialogue consists of 19 utterances and features a conversation between a female agent purchasing furniture for her office, and a male shop owner describing some furniture items that she could consider for her purposes. Results from a pilot study used for validation of the dialogue and the setting showed that the dialogue was acceptable to an English speaking audience (Breitfuss et al., 2009).

The dialogue was used as a template in which five first-mention REs could be varied. The REs used to fill out these slots were chosen carefully to cover various aspects of REs currently studied in the GRE literature. These aspects include: (1) cardinality, the REs targeted three singular objects and two larger sets of items; (2) locative expressions, the REs included three absolute locative expressions and two relative locative expressions; and (3) the position of the referent. The targets were distributed in the domain of conversation such that one referent was located near to the stationary agents, two referents were located far away from the agents, and two sets of referents were located somewhere in between those two extremes.

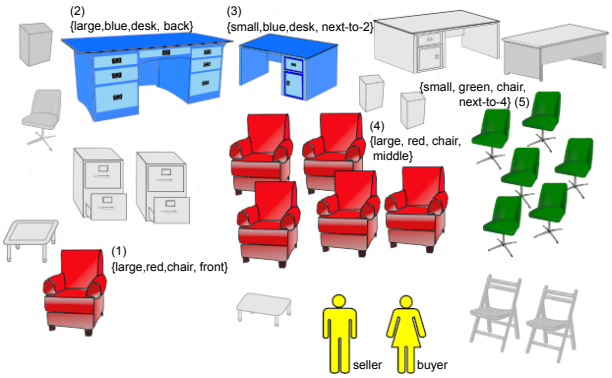


Figure 1: Bird's-eye sketch of the virtual furniture shop.

Figure 1 shows 14 furniture items that are used for assessing multimodal GRE output: (1) a large red chair (bottom left); (2) a large blue desk (top left), (3) a small blue desk (next to the large one); (4) a set of five large red chairs (in the middle), and (5) a set of six small green chairs (next to the set of reds), as well as a number of distractors (greyed-out items). We stipulated that the agents would stay stationary at the position indicated in Figure 1 and point in the direction where the targets can be found. The targets can be described with the attributes usually considered in GRE research (i.e. *type*, *colour*, *size*, *location*) and were realised as follows:

- RE1: large red chair in the front of the shop
- RE2: large blue desk in the back of the shop
- RE3: small blue desk next to it (where 'it' refers to the target of RE2)
- RE4: large red chairs in the middle of the shop
- RE5: small green chairs next to the red ones

The dialogue was translated to Japanese, Brazilian Portuguese and Dutch such that the dialogue was adapted to the normative, communicative and inferential rules of the respective cultures but the REs were as close to the English originals as possible. The translations and localisations for Portuguese and Dutch followed a similar pattern as the process for Japanese described in (Van der Sluis and Luz, 2011). Validation of the translated dialogues was conducted by three native speakers in the respective languages and revisions were made accordingly.

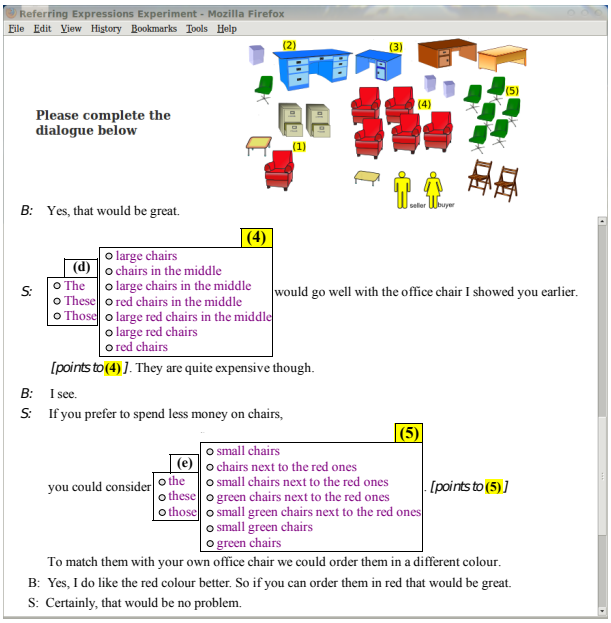


Figure 2: Screenshot of the application in which participants were asked to choose their preferred REs. Utterances by the Seller and Buyer are marked with "S:" and "B:", respectively. Options were presented as shown in the DE-boxes marked *(d)* and *(e)*, and RE-boxes marked *(4)* and *(5)*.

2.2 Materials

The study was conducted over the Web and consisted of three pages. The first page presented a tutorial in which the participants were told about the goals of the study, what they were going to see on the next page, and what they would be asked to do. The second page is shown in Figure 2. At the top of the screen a picture of the domain was presented. The bottom part of the screen contained the dialogue through which the participants could scroll and select the REs they preferred from a set of options, all of which were simultaneously available to the participant while reading the sentence. The picture of the domain was always visible on the top part of the screen. The five REs of interest were each presented with two boxes as illustrated by, for instance, the items marked (e) and (5) in Figure 2: the DE-box, in which participants could select a determiner or demonstrative and the RE-box, in which combinations of properties could be chosen.

The RE-box contained seven possible REs in which the inclusion of *colour*, *size* and *location* were varied; all REs contained the relevant value for *type* as a noun. For instance, in the case of RE2 the options would be 'large desk', 'blue desk', 'desk in the back', 'large blue desk', 'large desk in the back', 'blue desk in the back' and 'large blue desk in the back'. After each RE-box, it was indicated that the agent's utterance of the RE would be combined with a pointing gesture in the direction of the target. The DE-box offered a number of options to compose deictic expressions in line with the determiners available in the respective languages. We refer to (Luz and van der Sluis, 2011) for our analysis of the determiners that were collected with this study. The third page of our study consisted of a "thank you" note and information about a prize draw, as a reward for participating in the study. All materials used in this study were fully translated into the languages considered.

3 Hypotheses

Because we study the perception of REs by presenting them to potential users in their own language and localised contexts (i.e. a context adapted to the normative, communicative and inferential rules of their cultural background) we used null hypothesis significance testing. In other words, our null hypotheses are that participants do not differ in their preferences dependent on their cultural background. If significant differences are observed, we can regard these differences as evidence towards alternative hypotheses.

The hypotheses for the REs to be selected by the participants are based on findings from cognitive linguistics (Pechmann, 1989; Arts et al., 2010) that show that absolute properties (e.g. *colour*) are preferred over relative properties (e.g. *size*). Following Krahmer and Theune (2002) we expect locative expressions to be even less preferred than relative properties. In our set up we presented the discourse domain including the agents that featured in the dialogue in a two-dimensional fashion. However, we asked the participants to imagine that the furniture seller agent included a pointing gesture to accompany the linguistic descriptions to refer to the targets. Hence we asked participants to imagine the distinguishing effect that this pointing gesture would have in a three-dimensional environment. As we cannot be sure about the scope of these pointing gestures in

the minds of the participants and their effect on the distractor set on which the participants based their choice of RE, we decided to test two hypotheses, which are summarised in Table 1.

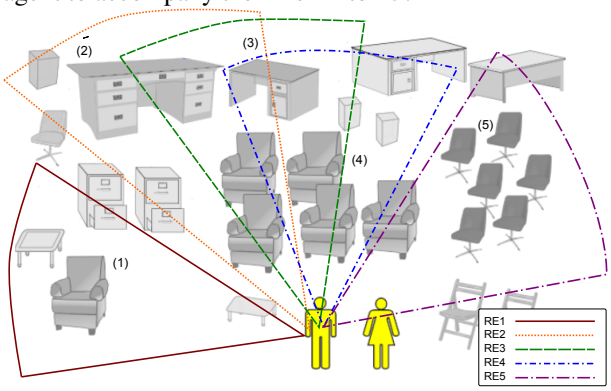
Table 1: Expected REs for referents *RE1* to *RE5* for two hypotheses *H1* and *H2* on the content of the REs.

Target	H1: Whole domain	H2: Gesture scope
RE1	colour, location	colour
RE2	colour, size	colour, size
RE3	colour, size	colour, size
RE4	colour, location	colour
RE5	colour, location	colour

Our first hypothesis, H1, is that participants in our study will consider all distractors in the domain as depicted in Figure 2 for each RE (i.e., the pointing gesture has no effect, it does not rule out any distractors). Accordingly, for RE1 we expect that participants will first include *colour* to rule out all objects in the domain that are not red. For RE1, *size* will not remove any distractors, but we expect that *location* will be included to rule out the group of red chairs in the middle of the shop. For RE2, we expect that *colour* will be selected to rule out all objects that are not blue. Secondly, *size* will be added to remove the remaining smaller blue desk and thereby empty the set of distractors. For RE3, we expect participants to include *colour* to rule out all distractors that are not blue and add *size* to rule out the large blue desk and thereby uniquely distinguish the target of RE3. RE4 will be distinguishing by first adding *colour* to rule out all objects that are not red. Then *location* will be added to remove the only remaining distractor, that is the singular red chair in the front of the shop. RE5 is expected to include *colour*, which leaves only green distractors, and *location* to remove the singular green chair on the left-hand side of the domain.

Our second hypothesis, H2, is that participants only consider the set of distractors located in the scope of the pointing gesture performed by the agent to distinguish the target. For all five targets we tentatively defined the scope of the pointing gestures as depicted in Figure 3, where the areas covered by the pointing gestures are of the same size, but differ in terms of the covered areas that include the target in the centre of the gesture's scope. Note, however, that the participants in our study were not provided with these gesture scopes, they had to imagine the effect of the gesture themselves. Accordingly, their representation might have been different from the scopes presented in Figure 3. For the sake of illustration, we define the set of distractors as including all objects that are located fully or partly within the projected lines that indicate the scope of the gesture. For all five REs we assume that the algorithm first adds a pointing gesture to the RE which results in a decrease of the number of distractors. For all five REs, however, inclusion of the pointing gestures does not result in distinguishing REs and participants are still expected to add linguistic properties to identify the targets uniquely. For RE1, *colour* should be added to empty the distractor set (i.e. the pointing gesture had already ruled out the group of red chairs in the middle of the shop). For RE2, *colour* and *size* are expected to be included; the pointing gesture's scope has decreased the target set but still includes some objects with a different colour as well as the smaller blue desk. RE3 also requires *colour* and *size* to respectively rule out the objects in the gesture's scope that are not blue as well as the large blue desk. Both RE4 and RE5 require *colour* to remove the remaining distractors located in the scope of the respective pointing gestures.

Figure 3: Furniture shop divided into five areas that cover the scope of the pointing gestures produced by the Seller agent to accompany the REs *R1* to *R5*.



3.1 Evaluation Metric

To test our hypotheses, H1 to H2, we compared the participants' choices with the realised output of a typical GRE algorithm that uses a preferred attribute

list alike the algorithm proposed by (Dale and Reiter, 1995) that mimics human preferences (i.e. [*colour*, *size*, *location*]). We chose the Dice coefficient as our evaluation metric, which accounts for a degree of overlap between two descriptions. Dice computes the degree of similarity between two sets by scaling the number of attributes that the two descriptions have in common, by the overall size of the two sets:

$$dice(H_a, R) = \frac{2 \times |H_a \cap R|}{|H_a| + |R|} \quad (1)$$

where H_a is the set of attributes in the description produced by a human author, and R the set of attributes in the reference description generated by the algorithm. Dice yields a value between 0 (no agreement) and 1 (perfect agreement). The attributes are chosen from a set $A = \{c, s, l\}$, denoting colour, size and location, respectively, so that possible H_a will be elements of $\mathcal{A} = 2^A \setminus \emptyset$. We summarise the Dice scores by their expected values for a particular object. That is, we report the mean scores weighted according to the probability p_a that a combination of attributes $a \in \mathcal{A}$ is chosen, as set out in equation (2).

$$E[dice(H, R)] = \sum_{a \in \mathcal{A}} p_a \times dice(H_a, R) \quad (2)$$

For comparison, we computed a baseline score (B) where p_a is a uniform distribution (i.e. all feature combination choices are equally likely) as a special case of (2) that is: $B = 1/7 \sum_{a \in \mathcal{A}} dice(H_a, R)$

The 'perfect recall percentage', (PRP), that is the proportion of times the hypotheses match the participants' choices exactly, is also reported.

4 Results

4.1 Participants

The address (URL) for the study was distributed through sending invitations for participation by email. Participants included 54 native speakers of Japanese (female: 26%(14), male: 74%(40)), 91 native speakers of English (female: 60%(55), male: 40%(36)), 42 native speakers of Brazilian Portuguese (female: 60%(25), male: 40% (17)) and 51 native speakers of Dutch (female: 55%(28), male: 45%(23)). Table 2 summarises the characteristics of the participants that took part in our study.

4.2 Referring Expressions

Table 3 presents the REs that were selected by the participants in our study per language group. As regards which RE was chosen by the majority of each language group we find that for RE1, 'large red chair in the front', speakers of Portuguese and Dutch agree in their selection of *colour* and *location*. In contrast, Japanese participants largely preferred the RE including only *colour* and English participants preferred to include all available properties in the description. For RE2, 'large blue desk in the back', a majority in all four language groups chooses to include all available properties. For RE3, 'small blue desk next to it', the majorities of the four language groups also agree and select a description that includes *size* and *location* (note that this is not a possible algorithmic output when we assume the proposed preference order in the current domain). However, for RE3, the Japanese data presents a tie, indicating that an equally large group of participants selected all available properties to distinguish the target. For RE4, 'large red chairs in the middle', Japanese and Portuguese speaking participants team up with a majority vote for inclusion of only *colour*, while both English and Dutch participants prefer *colour* and *location*. Finally, for RE5, 'small green chairs next to the red ones', the Japanese and Portuguese speakers again agree with a majority vote for *colour*, while Dutch participants select *colour* and *size* and English speakers prefer to include all available properties to refer to the target.

Per language group we find that the majority of the Japanese participants chose an RE that only include *colour* for RE1, RE4 and RE5 (all between 40 and 50%). For RE2 and RE3 the Japanese majority chose to include all available properties. The English participants show different preferences, namely including all available properties in RE1, RE2 and RE5, *size* and *location* for RE3, and for RE4 *colour* and *location*. Speakers of Portuguese and Dutch present more variability. Portuguese speakers select only *colour* for RE4 and RE5, while the majority prefers different descriptions for RE1, RE2 and RE3. The majority of Dutch speakers chooses *colour* and *location* for RE1 and RE4 and prefers various descriptions for RE2, RE3 and RE4.

Table 2: Participants in our study per *L*anguage (*E*nglish, *J*apanese, *P*ortuguese and *D*utch) in terms of *N*umber of subjects, number of subjects per *Age* band, where *1* = 20-30, *2* = 31-40, *3* = 41-50, *4* = 61-70 and *5* = over 70 years old, and per *Occupation* as *S*tudent, *A*cademic or *O*ther.

L	N	Age	Occupation
J	54	1=57%(31); 2=28%(15); 3=15%(8)	S=52%(28); A=13%(7); O=35%(19)
E	91	1=52%(47); 2=23%(21); 3=22%(20); 4=2%(2); 5=1%(1)	S=44%(40); A=26%(23); O=31%(28)
P	42	1=71%(30); 2=26%(11); 3=2 %(1)	S=29%(12); A=57%(24); O=%(6)
D	51	1=22%(11); 2=33%(17); 3=26%(13); 4=14%(7); 5=6%(3)	S=4%(2); A=14%(7); O=80%(42)

Table 3: Means and standard deviations of REs collected per *L*anguage (*E*nglish, *J*apanese, *P*ortuguese and *D*utch) for *RE1* to *RE5* for which the values of the available attributes *colour*, *size* and *location* are indicated, as well as the actual choices made by the participants in the study as combinations of *c*olour, *s*ize and *l*ocation. The PRP scores for H1 and H2 are presented in boldface.

L		**RE1**	**RE2**	**RE3**	**RE4**	**RE5**
	colour, size, location	*red, large, front*	*blue, large, back*	*blue, small, next*	*red, large, middle*	*green small, next*
J	c	**42.6%** (23)	7.4% (4)	3.7% (2)	**46.3%** (25)	**48.1%** (26)
E		**7.7%** (7)	0% (0)	0% (0)	**11.1%** (10)	**14.3%** (13)
P		**26.2%** (11)	2.4% (1)	2.4% (1)	**33.3%** (14)	**38.1%** (16)
D		**15.7%** (8)	2% (1)	0% (0)	**11.8%** (6)	**17.6%** (9)
J	s	7.4% (4)	14.8% (8)	9.3% (5)	3.7% (2)	9.3% (5)
E		1.1% (1)	1.1% (1)	4.4% (4)	1.1% (1)	4.4% (4)
P		4.8% (2)	0% (0)	0% (0)	2.4% (1)	4.8% (2)
D		3.9% (2)	2% (1)	9.8% (5)	2.0% (1)	0% (0)
J	l	1.9% (1)	3.7% (2)	5.6% (3)	1.9% (1)	0.0% (0)
E		3.3% (3)	3.3% (3)	4.4% (4)	0% (0)	2.2% (2)
P		0% (0)	0% (0)	14.3% (6)	4.8% (2)	7.1% (3)
D		5.9% (3)	3.9% (2)	9.8% (5)	9.8% (5)	3.9% (2)
J	cs	29.6% (16)	20.4% (11)	7.4% (4)	13% (7)	27.8% (15)
E		12.1% (11)	17.6% (16)	1.1% (1)	7.7% (7)	25.3% (23)
P		19% (8)	11.9% (5)	11.9% (5)	7.1% (3)	4.8% (2)
D		0% (0)	17.6% (9)	2% (1)	0% (0)	39.2% (20)
J	cl	**5.6%** (3)	**5.6%** (3)	**14.8%** (8)	**24.1%** (13)	**7.4%** (4)
E		**31.9%** (29)	**5.5%** (5)	**4.4%** (4)	**35.2%** (32)	**16.5%** (15)
P		**28.6%** (12)	**19%** (8)	**9.5%** (4)	**28.6%** (12)	**26.2%** (11)
D		**43.1%** (22)	**9.8%** (5)	**2%** (1)	**43.1%** (22)	**11.8%** (6)
J	sl	3.7% (2)	9.3% (5)	29.6% (16)	3.7% (2)	0.0% (0)
E		6.6% (6)	9.9% (9)	48.4% (44)	8.8% (8)	9.9% (9)
P		2.4% (1)	21.4% (9)	35.7% (15)	9.5% (4)	4.8% (2)
D		3.9% (2)	11.8% (6)	41.2% (21)	13.7% (7)	5.9% (3)
J	csl	9.3% (5)	38.9% (21)	29.6% (16)	7.4% (4)	7.4% (4)
E		37.4% (34)	62.6% (57)	37.4% (34)	25.3% (23)	27.5% (25)
P		11.9% (5)	45.2% (19)	26.2% (11)	14.3% (6)	14.3% (6)
D		27.5% (14)	52.9% (27)	35.3% (18)	19.6% (10)	21.6% (11)

4.3 Distractor Sets

Table 4 displays the Dice scores for the collected data and our baseline per hypotheses per language group computed for the REs for which the hypotheses rendered different output (i.e. RE1, RE4 and RE5). Recall that H1 predicts that participants would take all objects in the domain into account as distractors when selecting their preferred descrip-

tion, while H2 predicts that participants would only consider the objects located in the scope of the pointing gesture that would accompany the linguistic description. Except for the Japanese data for RE1 and RE5 on H1, all Dice scores seem well above the baseline. This reinforces that for all three REs the figures show that the choice of the speakers of Japanese matches H2 best, while the other three languages match better with H1. T-tests at the $p < .05$ level comparing the Dice scores per RE per language show significant differences for the collected English REs for the targets of all three REs (RE1 t=8.786, RE4 t=8.805 and RE5 t=3.574). For Japanese REs significant differences were found for the targets of RE1 and RE5 (RE1 t=3.046 and RE5 t=5.177). The Dutch data displayed significant differences for RE1 and RE4 (RE1 t=6.137 and RE4 t=8.058). Differences between the Dice scores for the Portuguese data are not significant.

Table 4: Dice scores for the RE1, RE4 and RE5 computed per *L*anguage (*E*nglish, *J*apanese, *P*ortuguese and *D*utch) and the *B*aseline, where significant differences between the Dice scores of H1 and H2 are denoted with '*' at the $p < .05$ level and '**' at the $p < .01$ level.

L		*H1-Dice*	*H2-Dice*	*H1 vs H2*
J	**RE1**	.59	.71	**
E		.78	.56	**
P		.71	.64	
D		.81	.58	**
B		.59	.40	
J	**RE4**	.70	.75	
E		.78	.52	**
P		.74	.64	
D		.80	.50	**
B		.59	.40	
J	**RE5**	.59	.75	**
E		.67	.56	**
P		.73	.66	
D		.66	.62	
B		.59	.40	

4.4 Cross-linguistic Findings

Table 5 displays the significant differences between the languages per hypotheses (H1: distractors = all objects in the domain safe the target, and H2: distractors = objects in the scope of the pointing gesture), per RE (RE1, RE4 and RE5) that were found through a multivariate ANOVA with posthoc Tukey's HSD tests. For RE1, 'large red chair in the front' the REs from the Japanese speakers significantly differed from all three other languages, indicating that the collected Brazilian Portuguese, English and Dutch REs better match H1 than the Japanese REs. For RE5 we also found a significant difference between the Japanese and the Portuguese group for H1. Results further show that for all REs the choices of the Japanese group differed significantly from the choices of the English group when comparing the Dice scores for hypothesis H2, indicating that H2 was a significantly better match for the REs selected by the participants in the Japanese group than the REs selected by the English group. For RE4, the Japanese REs also differed from the Dutch ones for H2.

Table 5: Multivariate ANOVA per referring expression (*RE1*, *RE4* and *RE5*), per hypothesis (*H1* and *H2*) reporting *Mean* differences and standard errors (*StdE*) for significant differences between language pairs (*E*nglish, *J*apanese, *P*ortuguese and *D*utch), where differences are denoted with '*' at the $p < .05$ level and '**' at the $p < .01$ level.

RE	*H*	*L-pair*	*Mean*(*StdE*)	*P*
RE1	H1	J - D	.22(.042)	**
		J - E	.19(.037)	**
		J - P	.12(.044)	*
	H2	J - E	.15(.048)	*
RE4	H2	J - D	.24(.062)	**
		J - E	.23(.045)	**
RE5	H1	J - P	.13(.045)	*
	H2	J - E	.20(.052)	**

Overall, we found significant effects between languages. RE1, large red chair in the front, showed such an effect for H1 (F(3,234)=11.903, *MSE*=.554 p <.001) and H2 (F(3,234)=3.482, *MSE*=.280 p <.05). RE4, 'large red chairs in the middle', only for H2 (F(3,234)=7.563, *MSE*=.280 p <.05), and RE5,'small green chairs next to the red ones', for H1 (F(3,234)=2.954, *MSE*=.143 p <.001) and H2 (F(3,234)=4.867, *MSE*=.438 p <.01).

5 Summary and Discussion

The REs collected with our web experiment display many differences between the four language groups

included in our study. Most notably is the fact that the majority of Japanese participants preferred shorter descriptions than the majorities of the participants in the other language groups. Especially, the Japanese majority chose only to include the property *colour* in the object descriptions RE1, RE4 and RE5, while the majorities of the English and Dutch participants also chose *location* and sometimes *size*. Interestingly, the Portuguese speakers, like the Japanese, chose only *colour* for RE4 and RE5.

For RE2, 'large blue desk in the back', the majorities of all four language groups agreed in selecting all available properties for the RE. This might be explained by the fact that the focus in the dialogue shifted from a furniture item in the front of the shop (i.e. the large red chair in the front located near to the agents) to the back of the shop (i.e. far away from the agents). Note that the target of RE3, 'small blue desk next to it' was equally far away from the agents as the target of RE2. However, when the target of RE3 is discussed in the script, the focus of attention was already in the back area of the shop.

As regards our hypotheses, we found that the REs selected by the Japanese participants best matched H2, indicating that they considered a reduced distractor set in composing their REs due to the scope of the accompanying pointing gesture. In contrast, the REs selected by the participants in the other language groups better matched H1, stating that people would consider all objects in the conversation domain as distractors when identifying targets.

We also found various significant differences between the Dice means of the four language groups per RE, indicating that Japanese speakers employ different strategies in composing REs than participants in the English, Dutch and Portuguese groups.

The fact that the Japanese participants in our study are predominantly male (74%) may have been a potential confounding factor in our results. As men are known to be less verbal than women, the reported effect could be a gender rather than a language effect. We ran a separate statistical analysis on gender effects on our Japanese data. It turned out that gender affected the use of the *location* attribute with t=3.05 at the $p < .05$ level indicating that Japanese females used *location* more often than Japanese males (in 57% and 36% of REs, respectively). Comparing the hypotheses H1 and H2 per object with respect to gender, Japanese males had a significant preference for H2 over H1 for RE1 (mean Dice scores 75% vs. 60%, t=2.38, $p < .05$) while Japanese females exhibited no clear preference (57% vs 58%, non-signif). Both genders preferred H2 for RE5 (56% vs 73% for males and 68% vs 82% for females, $p < .05$. There was no gender effect with respect to RE4. Further studies are required to investigate gender across different languages.

Another reason for the effects we observed in our study may be related to differences in the use of pointing gestures in the languages we considered. For instance, (Kita and Özürek, 2003) showed differences in gesturing between English and Japanese speakers (not about pointing though), and it is conceivable that the observed language differences are caused by gesture differences. In future work it would be interesting to add a condition to the experiment in which pointing gestures are not included.

6 Conclusion and Future Work

This paper has presented a cross-linguistic study of the production of REs by native speakers of English, Japanese, Dutch and Brazilian Portuguese, which displayed many significant differences between the language groups. These differences were related to the set of distractors that was taken into account, which was hypothesised to be influenced by the effect of pointing gestures that accompanied the REs. One limitation of this study is clearly that the pointing gestures to accompany the linguistic descriptions were scripted and the effect of those gestures in the minds of the participants could only be assumed. Instead of linguistically described pointing gestures, animations of pointing gestures may be more effective for deriving the effect of pointing on a linguistic description. We refer to (Van der Sluis et al., to appear) for an attempt in this direction.

Another limitation is that only five predefined realisations of REs were used to elicit object descriptions from the participants. The REs, however, were carefully chosen as to reflect on issues currently being studied in GRE. The situated and life-like dialogue that was used in the study, specially in terms of focus shifts, might also have influenced the participants' choice of REs. In addition, perhaps overhearer effects to do with attention and engagement

may have played a role. However, with our with 'static' study we have not attempted to mimic an interactive, real-time situation.

Upon completing their choices participants were offered the opportunity to enter free-form comments in a text box. From the participants' comments we know that people were positively engaged in the study. Some participants, however, indeed criticised the limited choice of descriptive attributes and their suitability for the sales domain. While the criticism is valid, our choice of REs was based on previous work on RE generation where the furniture domain is used very often (i.e., through the COCONUT corpus (Di Eugenio et al., 2000) and the TUNA corpus (Van Deemter et al., To Appear)).

In summary, although limited in terms of expressiveness, the range of attributes available allowed us to identify general differences in RE production styles between the languages. With inspecting almost 1200 REs, we can conclude that a typical GRE algorithm that uses a well established preference order does not match the human production of multimodal REs for all languages and further studies are necessary to inform the design of GRE algorithms that can be employed in multilingual, multimodal and interactive environments.

7 Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin.

References

E. André and T. Rist. 1996. Coping with temporal constraints in multimedia presentation planning. In *Proc. of the AAAI'96*.

A. Arts, A. Maes, L. Noordman, and C. Jansen. 2010. Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1):361–374.

W. Breitfuss, I. Van der Sluis, S. Luz, H. Prendinger, and M. Ishizuka. 2009. Evaluating an algorithm for the generation of multimodal referring expressions in a virtual world: A pilot study. In *Proc. of IVA-09*, Amsterdam. 2009.

R. Dale and E. Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263.

K. Van Deemter, A. Gatt, I. Van der Sluis, and R. Power. To Appear. Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*.

K. Van Deemter. 2002. Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52.

K. Van Deemter. 2006. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

B. Di Eugenio, P. Jordan, R. Thomason, and J. Moore. 2000. The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *Intl. Journ. Human-Comp. Studies*, 6:1017–1076.

K. Funakoshi, S. Watanabe, and T. Tokunaga. 2006. Group-based generation of referring expressions. In *Proc. of the INLG-06*, pages 73–80.

C. Gardent. 2002. Generating minimal definite descriptions. In *Proc. of the ACL-02*.

P. Jordan and M. Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

S. Kita and A. Özürek. 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48(1):16–32.

E. Krahmer and M. Theune. 2002. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. CSLI Publications.

E. Krahmer, S van Erk, and A. Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth. 2006. Deictic object reference in task-oriented dialogue. In G. Rickheit and I. Wachsmuth, editors, *Situated Communication*. Mouton de Gruyter.

S. Luz and I. van der Sluis. 2011. Production of demonstratives in Dutch, English and Portuguese dialogues. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG'11)*.

D. Pareira and I. Paraboni. 2008. From TUNA attribute sets to Portuguese text: a first report. In *Procs. of INLG'08*.

T. Pechmann. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27:89–110.

P. Piwek. 2009. Salience and pointing in multimodal reference. In *Proc. of preCogsci 2009: Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference. At CogSci'09*, Amsterdam, The Netherlands.

I. Van der Sluis and E. Krahmer. 2007. Generating multimodal referring expressions. *Discourse Processes*, 44(3):145–174.

I. Van der Sluis and S. Luz. 2011. Issues in translating and producing Japanese referring expressions for dialogues. *Linguistic Issues in Language Technology*, 5(1):1–46.

I. Van der Sluis, S. Luz W. Breitfuß, M. Ishizuka, and H. Prendinger. to appear. Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world. *International Journal of Human-Computer Studies*.

P. Spanger, Y. Masaaki, I. Ryu, and T. Takenobu. 2009. A Japanese corpus of referring expressions used in a situated collaboration task. In *Proc. of the ENLG-09*.

M. Theune, R. Koolen, and E. Krahmer. 2010. Cross-linguistic attribute selection for REG: Comparing Dutch and English. In *Procs. INLG'10*.