Lang Resources & Evaluation (2013) 47:1285–1304
DOI 10.1007/s10579-013-9240-5

ORIGINAL PAPER

A task-performance evaluation of referring expressions in situated collaborative task dialogues

Philipp Spanger · Ryu Iida · Takenobu Tokunaga · Asuka Terai · Naoko Kuriyama

Published online: 21 June 2013

Abstract Appropriate evaluation of referring expressions is critical for the design of systems that can effectively collaborate with humans. A widely used method is to simply evaluate the degree to which an algorithm can reproduce the same expressions as those in previously collected corpora. Several researchers, however, have noted the need of a task-performance evaluation measuring the effectiveness of a referring expression in the achievement of a given task goal. This is particularly important in collaborative situated dialogues. Using referring expressions used by six pairs of Japanese speakers collaboratively solving Tangram puzzles, we conducted a task-performance evaluation of referring expressions with 36 human evaluators. Particularly we focused on the evaluation of demonstrative pronouns generated by a machine learning-based algorithm. Comparing the results of this task-performance evaluation with the results of a previously conducted corpus-matching evaluation (Spanger et al. in Lang Resour Eval, 2010b), we confirmed the limitation of a corpus-matching evaluation and discuss the need for a task-performance evaluation.

Keywords Task-performance evaluation · Referring expressions · Demonstrative pronouns · Situated dialogue · Japanese

1 Introduction

Automatic generation of effective referring expressions—as a linguistic means referring to a specific object in a particular situation—is critical in order to ensure

P. Spanger · R. Iida · T. Tokunaga (✉)
Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan
e-mail: take@cl.cs.titech.ac.jp

A. Terai · N. Kuriyama
Department of Human System Science, Tokyo Institute of Technology, Tokyo, Japan

smooth collaboration between humans and computers. Early research on the generation of referring expressions focused on generating isolated expressions in a static environment such as expressions referring to a referent given a set of objects with various attributes (Dale 1989; Dale and Reiter 1995). However, such static environments do not reflect actual environments where humans collaboratively perform a given task through interactions. In fact, the importance of human reference behaviour in dynamic environments has been noted very early on Bolt (1980), Clark and Wilkes-Gibbs (1986), Heeman and Hirst (1995).

With the development of increasingly complex generation algorithms, there has been a heightened interest in and an ongoing significant discussion about various evaluation measures for generated referring expressions (Reiter and Belz 2009). At bottom, there are two different approaches to evaluation: *intrinsic* and *extrinsic* methods (Sparck Jones and Galliers 1996). Intrinsic methods often measure similarity between the system output and a gold standard corpus by using metrics such as tree similarity, string-edit-distance, BLEU (Papineni et al. 2002). Intrinsic methods have recently been popular in the Natural Language Generation (NLG) community.

On the other hand, extrinsic methods evaluate the system output based on an external metric, such as its impact on human task performance. From early on in the NLG community, extrinsic task-performance evaluation has been considered as the most meaningful evaluation, especially for convincing people in other communities of the usefulness of a system (Reiter and Belz 2009). Task-performance evaluation is recognised as the "only known way to measure the effectiveness of NLG systems with real users" (Reiter et al. 2003). Following this direction, task-performance evaluation has been utilised in some areas in NLG. For example, in the GIVE challenges (Byron et al. 2009; Koller et al. 2010), automatically generated instructions to video game players are evaluated in terms of their success in achieving the game goal.

A series of different extrinsic evaluation metrics has been proposed for generated linguistic expressions considering the purpose of the expressions, such as the number of mistakes caused by generated instructions (Young 1999), the persuasive power of generated argumentative text (Carenini and Moore 2006), the learning gain of generated pedagogical text (Di Eugenio et al. 2002). Referring expressions are generally used with a clear intention to allow the hearer to identify the target. Hence, in the implementation of a task-performance evaluation for referring expressions, we can utilise the performance of human evaluators in identifying the target as an evaluation metric.

The realisation of an extrinsic evaluation for referring expressions in a dynamic environment faces one key difficulty. In a static environment, an extrinsic evaluation can be realised relatively easily by having evaluators identify the target by providing a *static* context (e.g. a still image) and a referring expression, even though this is still costly in comparison to intrinsic methods (Belz and Gatt 2008; Belz et al. 2010). In contrast, an extrinsic evaluation in a *dynamic* environment needs to present evaluators with a *dynamic* context (e.g. a certain length of the dialogue fragment) preceding a referring expression, before asking the evaluators its referent.

It is clearly not feasible to simply show the *whole* preceding dialogue; this would make even a very small-scale evaluation too costly. Thus, in order to realise a cost-effective extrinsic evaluation in dynamic environments, we need to carefully consider the preceding context to be presented to evaluators. We have already discussed this issue elsewhere (Spanger et al. 2010a), and use the results in this research.

In this paper, we conduct an extrinsic task-performance evaluation of generated Japanese demonstrative pronouns to compare the results with that of an intrinsic corpus-matching evaluation conducted in our previous work (Spanger et al. 2010b). Our results support the past research claim that intrinsic and extrinsic evaluations focus on different aspects of generated referring expressions (Belz and Gatt 2008; Belz et al. 2010).

This paper is organised as follows. Section 2 provides an overview of our previous work on the evaluation of referring expressions and motivates the need to work towards the implementation of extrinsic task-performance evaluation, particularly in dynamic environments. Section 3 describes the characteristics of the REX-J corpus utilised in our experiments. Section 4 briefly explains a generation model of demonstrative pronouns (DPs) proposed in our previous work (Spanger et al. 2010b), which adopts a machine learning technique to classify a pair of a target object and a situation into two cases: DP use and non-DP use. This section also provides the intrinsic evaluation results of the model. Section 5 discusses the implementation of an extrinsic task-performance evaluation following the discussion in Sect. 4. Section 6 presents the conclusions as well as future work.

2 Related work

As laid out in Sect. 1, an evaluation method of referring expressions can be categorised according to whether it is intrinsic or extrinsic. In addition, the environment in which referring expressions are used can be categorised according to whether it is static or dynamic. We review previous work in terms of these two aspects.

Intrinsic and extrinsic evaluations are based on different assumptions. Namely, intrinsic evaluations assume that there are ideal expressions in a given context, and thus the system performance can be measured by the degree to which the system can reproduce them. Expressions in a corpus are often used as such ideal expressions. On the other hand, extrinsic evaluations do not assume ideal expressions and evaluate the system in terms of some external task performance, for instance, how correctly humans can identify the referent of system-generated referring expressions.

Intrinsic evaluation has been widely used in NLG (e.g. Lester et al. 1999; Reiter et al. 2005; Cahill and van Genabith 2006) and the recent competitive 2009 TUNA challenge (Gatt and Belz 2010), since it enables rapid and automated evaluations of developing systems. It is also suitable when the system's explicit goal is to model or reproduce human performance. However, several researchers have pointed out its limitation. Reiter and Sripada (2002) argue, for example, that a generated text might

be very different from those in a reference corpus but still might achieve the specific communicative goal.

Corpus-similarity metrics measure how well a system reproduces what speakers (or writers) did, while another important consideration is its effect on humans (i.e. hearers or readers). For instance, Khan et al. (2009) argue that "most existing evaluations are essentially speaker-oriented ... disregarding their effectiveness." Furthermore, there are experimental results showing that intrinsic and extrinsic evaluation measures are not significantly correlated (Belz and Gatt 2008; Belz et al. 2010). Such discrepancy might come from the "egocentricity" of speakers, i.e. speakers do not always try to use optimal expressions for hearers (Horton and Keysar 1996). In summary, intrinsic and extrinsic evaluation measures look at the different aspects of referring expressions, thus both should be taken into account for system evaluations (Belz and Gatt 2008; Belz et al. 2010; Krahmer and van Deemter 2012).

The difference between static and dynamic environments is characterised by whether a situation in which referring expressions are used can change. A static environment is one such as the TUNA corpus (van Deemter et al. 2012), which collected referring expressions based on still images. In contrast, a dynamic environment comprises a constantly changing situation where humans need context information to identify the referent of a referring expression.

Referring expressions in the static environment have been evaluated relatively extensively. van der Sluis et al. (2007) conducted an intrinsic evaluation, measuring corpus-similarity in terms of the Dice coefficient. The series of the GRE (Generating Referring Expressions) Challenges systematically evaluated various systems mainly focusing on intrinsic evaluation in static environments (Belz and Gatt 2008; Belz et al. 2010; Belz and Kow 2010). More recently, van Deemter et al. (2012) provided an extensive discussion on an intrinsic evaluation using the TUNA corpus.

There have been extrinsic evaluations in a static environment as well, such as Paraboni et al. (2007) and Khan et al. (2009). Paraboni et al. (2007) evaluated to what extent redundant information impacts on the readers' referent identification performance. Khan et al. (2009) evaluated generated referring expressions in terms of readers' processing speed, reading speed and comprehension speed.

Over the recent period, while research on the generation of referring expressions has moved to dynamic environments such as situated dialogue (e.g. Jordan and Walker 2005; Stoia et al. 2006), the preponderance of intrinsic evaluation has continued. Both Jordan and Walker (2005) and Stoia et al. (2006) carried out intrinsic evaluations measuring corpus-similarity or asking evaluators to compare system outputs with human generated expressions. Gupta and Stent (2005) also conducted an intrinsic evaluation using dialogue corpora: the Map Task corpus (Anderson et al. 1991) and the COCONUT corpus (Di Eugenio et al. 2000).

Table 1 summarises the categorisation of the previous work we mentioned above as well as the position of our research. The rows of Table 1 correspond to the contrast of intrinsic and extrinsic evaluation, while the columns represent the characteristics of environments in which referring expressions are used.

Table 1 Overview of recent work on evaluation of referring expressions

Method	Environment	
	Static	Dynamic
Extrinsic	Paraboni et al. (2007) Belz and Gatt (2008) Khan et al. (2009) Belz et al. (2010)	Foster et al. (2009) Campana et al. (2011) The present work
Intrinsic	van der Sluis and Krahmer (2007) Belz and Gatt (2008) Belz and Kow (2010) Belz et al. (2010) van Deemter et al. (2012)	Jordan and Walker (2005) Gupta and Stent (2005) Stoia et al. (2006) Foster et al. (2009)

As one can see from Table 1, while in the static environment both intrinsic and extrinsic evaluations have been considered, little attention has been paid to the extrinsic evaluation of referring expressions in the dynamic environment. Foster et al. (2009) evaluated generation strategies of instructions for a collaborative construction task by both considering intrinsic[1] as well as extrinsic measures. Unlike most intrinsic evaluation research where individual referring expressions were evaluated, their main concern is evaluating generation strategies, particularly the interaction between the text structure and usage of referring expressions. Therefore, their work aims at a macroscopic evaluation rather than the evaluation of individual referring expressions. Campana et al. (2011) employed the dual-task methodology[2] for measuring humans' cognitive load in interpreting referring expressions. Carefully designed referring expressions under various conditions were interpreted by human subjects in the experiments. They found that human-like referring expressions required less cognitive load, although there was no significant difference in the accuracy of referent identification.

Recently, extrinsic evaluation of automatically generated instructions in a situated dialogue has been actively studied in the GIVE and GIVE-2 Challenge framework (Byron et al. 2009; Koller et al. 2010; Striegnitz et al. 2011). Gargett et al. (2010) analysed the GIVE-2 corpus which was collected through keyboard-based interactions in a treasure hunting video game. The evaluation measures used in their analysis are macroscopic such as interaction time, the number of object manipulations and the number of instructions. Their extrinsic evaluation does not include the effectiveness of referring expressions yet. Unlike these previous studies, we tackle the design and implementation of an extrinsic evaluation of individual referring expressions in a situated dialogue setting (the upper-right cell in Table 1).

[1] They compared the system output with subjective human judgement as a gold standard rather than an existing corpus.

[2] In this methodology subjects are requested to perform different kinds of tasks simultaneously, one of which is the target task to measure the cognitive load.

As mentioned in the previous section, extrinsic evaluation of referring expressions in dynamic environments requires determining a necessary context. In order to investigate the context necessary to understand a referring expression in a situated collaborative dialogue, we have carried out an experiment with 33 evaluators in the same domain as we deal with in this research (Spanger et al. 2010a). On the basis of the corpus analysis of the domain in question, we focused on two events prior to a referring expression, namely the last mention to the intended target object and the last action on the target. The former importance has often been mentioned in the discourse research community, particularly in anaphora resolution research (Mitkov 2002), and the latter is important in situated collaborative dialogues like our current domain (Spanger et al. 2010b). The conclusion of our previous work (Spanger et al. 2010a) is that including both events in a preceding context before a referring expression provides a reasonable context for hearers to identify its referent. Based on this conclusion, we design the experiment for our extrinsic task-performance evaluation.

3 Referring expressions in the REX-J corpus

For the research discussed in this paper, we utilise the REX-J corpus (Spanger et al. 2010b), a Japanese corpus of referring expressions in a situated collaborative task.[3] It was collected by recording the interaction of pairs of dialogue participants solving the Tangram puzzle cooperatively. The goal of the Tangram puzzle is to construct a goal shape by arranging seven pieces of simple figures as shown in Fig. 1.

In order to record the precise position of every piece and every action by the participants, we implemented a Tangram simulator. The simulator displays two areas as shown in Fig. 1: a goal shape, and a working area where pieces are shown and can be manipulated. We assigned different roles to the two participants of a pair: *solver* and *operator*. The solver can see the goal shape but cannot manipulate the pieces and hence gives instructions to the operator; by contrast, the operator can manipulate the pieces but cannot see the goal shape. The solver and operator sit side by side and collaboratively solve the puzzle through dialogue. Each participant has her/his own computer display sharing the working area where the movement of the mouse cursor and the pieces is shown in real-time. A shield screen was set between them to prevent the operator from seeing the goal shape on the solver's display.

The REX-J corpus contains a total of 1,443 tokens of referring expressions in 24 dialogues (4 dialogues of 6 pairs) with a total time of about 4 h and 17 min. The average length of each dialogue is 10 min 43 s. The asymmetric role assignment encouraged referring expressions from the solver (solver: 1,243 tokens, operator: 200 tokens). We exclude from consideration 201 expressions referring to either groups of pieces or whose referent cannot be determined due to ambiguity, thus leaving us 1,242 expressions referring to a single unique piece.

[3] The corpus is publicly available together with other variants as the REX corpora (Tokunaga et al. 2012) through GSK (http://www.gsk.or.jp/index_e.html) (Resource ID: GSK2013-A).

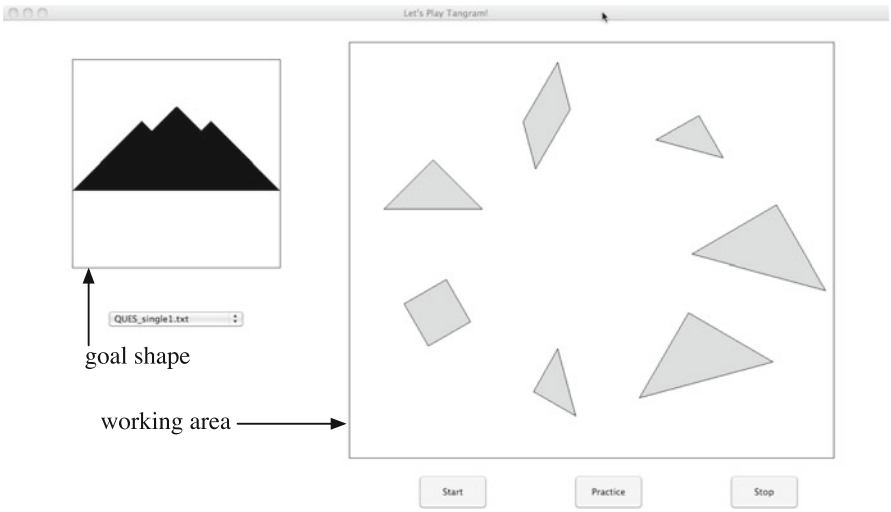

Fig. 1 Screenshot of the Tangram simulator

Table 2 Syntactic and semantic features of referring expressions in the REX-J corpus

Feature	Tokens	Example
(a) Demonstratives	742	*sore* (that)
(b) Attributes	795	*tittyai sankakkei* (the small triangle)
(c) Spatial relations	147	*hidari no okkii sankakkei* (the small triangle on the left)
(d) Action-mentioning	85	*migi ue ni doketa sankakkei* (the triangle you put away to the top right)

We identified syntactic/semantic features in the collected referring expressions as listed in Table 2: (a) demonstratives (adjectives and pronouns),[4] (b) object attributes, (c) spatial relations and (d) actions on an object. The underlined part of the examples denotes the feature in question. Spanger et al. (2010b) describes the further details of the corpus.

4 Intrinsic evaluation of generated demonstrative pronouns

In this section, we review an intrinsic evaluation of generated referring expressions in our previous work (Spanger et al. 2009, 2010b), and provide a detailed error

[4] There are three types of demonstrative pronoun/adjective in Japanese: "*kore/kono* (this)", "*sore/sono* (that)" and "*are/ano* (that)". They are basically chosen based on the physical and mental distance between the speaker and the target (Ono 1994).

Table 3 Features describing a situation

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

analysis of the result. Based on this analysis, we discuss the insufficiency of this intrinsic evaluation and claim the need for a task-performance evaluation.

We particularly focus on evaluating generated demonstrative pronouns, since demonstrative pronouns are one of the most common referring expressions in a number of different domains and tend to be influenced by extra-linguistic information. As shown in Table 2, demonstratives are the dominant type of referring expressions in our corpus; we have 548 instances of demonstrative pronouns among 742 demonstratives.

4.1 Evaluation setup and results

In order to replicate human usage of demonstrative pronouns in the corpus, we employed a machine-learning approach utilising extra-linguistic as well as linguistic information, which is represented as features in a Support Vector Machine (SVM) (Vapnik 1998). Table 3 shows a list of the features we used for describing a situation in which a referring expression is generated. The features are categorised into three categories: the dialogue history features (D1–D5), the action history features (A1–A5) and the current operation features (O1 and O2). The dialogue features model the dialogue history, while the operation and action features capture key aspects of the operations to achieve the task goal.

We constructed an SVM classifier which classifies a pair comprised of a target and a situation represented by the features in Table 3 into two classes: "demonstrative pronoun (DP)" and "other (non-DP)", which mean "to generate a DP" and "to generate an expression other than DP", respectively. We evaluated the performance of the classifier by the extent to which the classifier correctly predicts the usage of DPs. The correctness is judged by comparing the classifier output with the referring expression actually used in the corpus. In our experiments, we utilised the SVM-light software[5] (Joachims 1999) with 1,242 instances of referring expressions extracted from our corpus. A linear kernel was adopted with a

[5] http://svmlight.joachims.org/.

Table 4 Results of classification

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

Table 5 Confusion matrix of the classification result

Corpus\system	DP	Non-DP
DP	421 (TP)	145 (FN)
Non-DP	133 (FP)	543 (TN)

default value for the parameter c. Since the size of our data is small, we conducted a tenfold cross validation.

Table 4 shows the overall results of the classification. The baseline algorithm ("Baseline") suggests to use a DP, whenever the most recently mentioned object is the target, and suggests not to use a DP otherwise. "All features" denotes the classifier using all features shown in Table 3. The next three rows denote each classifier removing the features in one of the feature categories: dialogue history, action history and current operation. We see that using all features performs better than the baseline in terms of F-measure. In contrast, removing the current operation features significantly degrades the performance, which is worse than the baseline. This suggests that a deictic usage of demonstrative pronouns is dominant in our corpus.

4.2 Error analysis

In order to conduct a detailed error analysis, we made a confusion matrix of the classification result when using all features (row "All features" in Table 4) as shown in Table 5. We need to distinguish two types of errors: *false positives* (FPs: when humans did not use DPs but the classifier suggested it) and the opposite case of *false negatives* (FNs: when humans used a DP but the classifier did not suggest it).

In the above intrinsic evaluation, the classifier output should strictly match the corpus expression in a given situation. Spanger et al. (2009) manually investigated the worst 50 FP instances in terms of the SVM outputs and noted that DP in those instances would certainly be acceptable and enable hearers to identify the target. This leads to the question whether using DPs as suggested by the classifier instead of the non-DPs that appeared in the corpus really makes it more difficult for humans to identify the target. This is the main motivation of conducting an extrinsic task-performance evaluation, which will be discussed in the next section.

5 Task-performance evaluation of generated referring expressions

This section describes an extrinsic task-performance evaluation of our classifier, particularly focusing on the FP instances of the intrinsic evaluation discussed in the previous section. The evaluation is conducted by presenting human evaluators with a fragment of dialogue in which a referring expression is replaced by a DP when the classifier suggested it. We then measure the performance of referent identification of the evaluators. To our knowledge, there has not been any previous work presenting such a task-performance evaluation of generated referring expressions in a situated dialogue. As mentioned in Sect. 2, Campana et al. (2011) made a similar attempt using carefully designed short discourse contexts rather than natural dialogue contexts.

Our classifier does not generate any alternative expression which would be necessary to provide to the evaluators when it predicts a non-DP. Hence, we were not able to implement a task-performance evaluation of the FN instances. We compare the referent identification performance of humans between the TP (true positive) instances and the FN instances, expecting that the performance of the FN instances will be worse than that of the TP instances. If this is the case, we can assume that humans have difficulty in resolving the referent of DPs in the situation corresponding to the FN instances, thus using DPs in such situations is not preferable for hearers. This would give an indirect evidence of that the classifier's output (suggesting not using DPs) is acceptable.

5.1 Experimental design

Figure 2 shows a screenshot of the interface presented to human evaluators. In a trial, the evaluator clicks the button (D) to start a video that captured the shared working area of the Tangram simulator. The audio of the dialogue fragment is provided in synchronisation with the video. The video and audio stop just before the utterance including a referring expression for evaluation. That next utterance is shown as text in the area (C), highlighting the referring expression in red colour. Considering the preceding dialogue context, the evaluator selects one of the numbered puzzle pieces as its referent by clicking a corresponding numbered button in the area (E). When the evaluator cannot identify a referent, she/he can click the "I don't know" button (the right-most one). The evaluator can replay the video as many times as she/he likes by clicking the button (D) before selecting a referent. In order to keep consistency of the evaluator's role, we utilise only referring expressions used by solvers for the evaluation. This leads to the evaluator consistently playing the operator role.

In order to implement this experiment, we need to decide the length of dialogue fragment presented prior to the utterance including the referring expression. This is unlike the task-performance evaluation of referring expressions in a static environment like the TUNA Challenge (Gatt and Belz 2010) where such a context is not necessary. As mentioned in Sect. 2, we have already investigated the appropriate length of the preceding dialogue context in the same Tangram domain

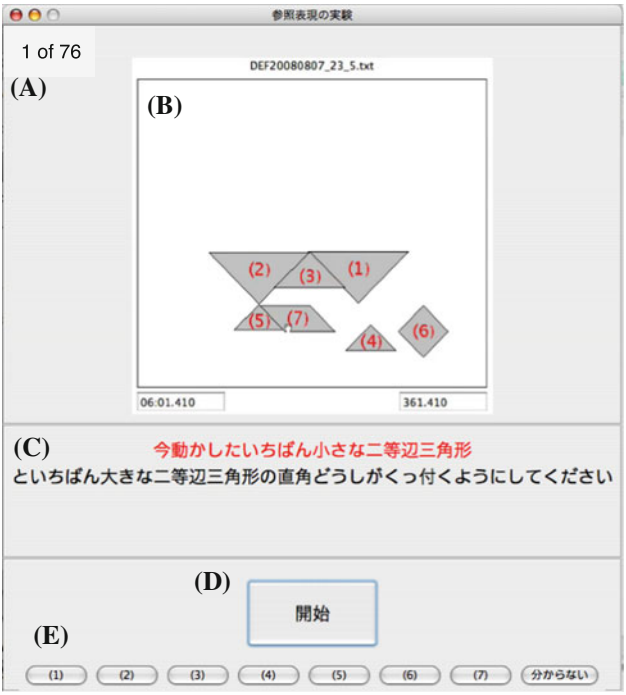

Fig. 2 Interface presented to evaluators

Table 6 Number of instances for the evaluation experiments

Instances\error type	TP	FN	FP
In the entire corpus	421	145	133
With the necessary context	191	64	36
Selected for evaluation	20	20	36

through experiments (Spanger et al. 2010a). Following the results of our previous work, we decided to adopt the duration of the preceding context in which both events, the last mention to the target and the last action on the target, are included. In addition, we limited the maximum duration to 30 s for practical reasons.

Among the referring expressions in the entire corpus, some lack the dialogue context satisfying the above conditions, e.g. an expression in the first utterance in a dialogue. The second row of Table 6 shows the number of instances that have a dialogue context satisfying the above conditions. Since extrinsic task-performance evaluation of the FP cases is our primary motivation for this experiment, we use all 36 instances for the evaluation. From the experiment in our previous work (Spanger et al. 2010a), we concluded that about 80 trials constitute a rough "maximum" limit of cognitive load for an evaluator. In addition to 36 instances of the FP cases, we utilised each 20 instances of the TP and FN cases. Given that the number of the usable TP and FN instances is much larger than 20, we have to decide on a basis for

selection. Spanger et al. (2010a) investigated the impact of context and emphasised the importance of the mouse cursor information in the generation of demonstrative pronouns. Hence, we decided to account for the distribution of the mouse cursor information in the corpus and seek to replicate this distribution as much as possible in the selected TP and FN instances. In our framework, the information of the mouse is represented in the form of the current operation features O1 and O2 (Table 3). We chose samples by keeping the distribution of selected instances with those features roughly the same as that of the entire corpus.

Next, we split the 36 FP instances in halves into two sets: FP_1 and FP_2. Of these 36 instances, there were 15 expressions of "*heikousihenkei* (the parallelogram)". Since there is only one parallelogram among the puzzle pieces, this expression is by itself a unique identifier without ambiguity. We tried to distribute these instances to the two sets as evenly as possible; 7 in FP_1 and 8 in FP_2. Then we derived a modified test set from each of these sets by replacing the referring expression with a Japanese demonstrative pronoun "*sore* (that)".[6] We call the original test set "FP-Orig" and the modified one "FP-DP" hereafter.

We recruited 36 Japanese native speakers (2 female and 34 male university students) as evaluators and split them into two groups. Each group was assigned one of two combinations of the FP instances: FP_1-Orig + FP_2-DP and FP_1-DP + FP_2-Orig. This set-up enables us to compare the effect of opposing expressions, the original expressions in the corpus (generated by humans) and demonstrative pronouns (suggested by the classifier), on referent identification in the same situation.

In addition to the 36 FP instances, both groups were assigned the same 20 TP and 20 FN instances, in which the original expressions in the corpus were presented. In summary, Each evaluator was assigned a trial set composed of the 76 trials (36 FPs, 20 TPs, 20 FNs). These trials were randomised in order when presented to the evaluators. The average time of a video clip was 15.2 s ($SD = 7.9$ s).

5.2 Evaluation metrics

In a task-performance evaluation, we need to decide how to measure "human performance". In this research, given the clear objective of referring expressions in task-oriented dialogue, i.e. referent identification, we evaluate the performance by measuring how effectively and efficiently a specific referring expression enables human evaluators to correctly identify the target. We utilise the following three measures:

1. Identification Rate (IR)

$$\mathrm{IR} = \frac{\text{the number of correct responses}}{\text{the number of responses}},$$

[6] In our current setting, "*sore/sono* (that)" is the most appropriate because the solver (speaker) does not have control of the mouse, i.e. pointing device. Actually, "*sore* (that)" is the most dominant demonstrative pronoun in the entire corpus.

2. Majority Identification Rate (MIR)

$$\text{MIR} = \frac{\text{the number of correctly identified test instances by a majority of evaluators}}{\text{the number of test instances}},$$

3. Identification Time (IT)

$$\text{IT} = \text{time of selecting a target by a evaluator.}$$

We note that at bottom IR and MIR are two measures that supplement each other. Together they characterise the effectiveness of referring expressions in terms of the success rate in identifying their referents. MIR indicates whether a majority of evaluators was able to identify the target in a specific context. This provides a different perspective on identification accuracy from IR, which only characterises responses by a single evaluator.

While IR and MIR indicate the effectiveness of referring expression, IT indicates the efficiency of identifying the referent. Obviously, an expression that allows faster target identification should be considered better as we can assume that it requires a lower cognitive load. We define IT as starting from the onset of displaying a solver's utterance to the evaluator's selection of a referent. Hence IT includes the reading time of the utterance as well as the identification time (including time for repeat views of the video).

IR and IT are widely used in this research field. Our definition of IT is, however, different from other work such as (Belz and Gatt 2008), which separately measured (1) the reading time (the time necessary to read a referring expression) and (2) the identification time (the time necessary to identify the target after reading the referring expression). Our IT definition combines both time periods together. In actual human communication, reading (or hearing) a referring expression and identifying its referent are not fully separable. Furthermore, separate measurement also ignores the fact that people often halfway through an expression understand its referent and act on this understanding. In our view, our definition of IT (reading and identification time) seems appropriate in our problem setting. Given that both approaches have their respective advantages, we would suggest that neither is superior.

5.3 Results and discussion

Among the data from 36 evaluators, we discarded two of them due to a system failure in the middle of a trial sequence, leaving us with the data from 34 evaluators. Luckily, they were from different evaluator groups, and thus we have data of 17 evaluators for each different test set. Given that each evaluator solved 76 trials, we have $34 \times 76 = 2{,}584$ responses to analyse. During the experiment, the evaluators were advised to alert the experiment instructors whenever they mistakenly selected a target due to a wrong operation. This occurred in 3 cases and we manually corrected the automatically collected data according to the evaluator's declaration. These manually corrected cases were excluded for the calculation of IT. Table 7 shows the identification accuracy by the evaluators.

Table 7 Referent identification accuracy (IR, MIR)

	TP	FN	FP-DP	FP-Orig
IR	0.972 (661/680)	0.725 (493/680)	0.894 (547/612)	0.883 (541/612)
MIR	1.000 (20/20)	0.800 (16/20)	0.944 (34/36)	0.889 (32/36)

The first two rows of Table 7 show the micro-averaged IR as well as the number of correct responses over all responses by the evaluators. However, analysing solely the IR values for all evaluators and all responses does not give us any clue as to what particular test instances were difficult for the evaluators. In order to capture the difficulty of referent identification of each test instance, we added a row for the MIR values together with the number of instances for which the majority of the evaluators correctly identified the referent.

We analyse the results of the evaluation experiment in terms of two aspects. The first is a comparison of the FP instances presented to the evaluators just as they appeared in the corpus (FP-Orig) with the same instances where the original expression was replaced by a DP (FP-DP). This analysis corresponds to a comparison of the two right-most columns of Table 7.

The second is a comparison of the TP instances with the FN instances. This analysis in turn corresponds to a comparison of the two left-most columns.

5.3.1 Comparison of the FP-DP and FP-Orig instances

Since we have responses to the FP-DP and FP-Orig instances in the same dialogue context, these responses can be directly compared. Interestingly, Table 7 shows that the IR measures of FP-DP and FP-Orig are almost the same, with FP-DP having a slightly higher IR of about 1 %. In the MIR measure, the identification accuracy for FP-DP is almost 5 % higher than FP-Orig. We can conclude from this result that replacing the referring expressions used in the corpus with a DP as our classifier suggested yields no decrease in the IR and MIR measures. The DPs in the same context perform as effectively as the expressions originally used by the dialogue participants in the corpus.

We also investigated the "agreement score", i.e. the ratio of agreement among evaluators on the majority selection. This score was calculated based on the instances where the majority of evaluators selected the correct target. We define the "agreement score" as the ratio of the number of evaluators agreeing on the correct target to the number of all evaluators (34 evaluators for TPs and FNs and 17 evaluators for FP-DPs and FP-Orig). Interestingly, the agreement score for FP-DP is 0.926 ($SD = 0.108$) while it is 0.969 ($SD = 0.088$) for FP-Orig. Hence, when correctly selecting the target, there is a slightly higher agreement between evaluators for FP-Orig than for FP-DP.

The number of instances for which a majority of evaluators failed to identify the target was 2 for the FP-DP instances and 4 for the FP-Orig instances. Among these 4

FP-Orig instances, there were 3 cases which mentioned some action on the target piece, e.g. "*ima ugokasita sankakkei* (the triangle you just moved)", which we named *action-mentioning expression (AME)* in our previous work (Spanger et al. 2010b). The AMEs are quite common in our current corpus because the goal of the task, solving the Tangram puzzle, necessarily requires various actions on the puzzle pieces. It is natural for the dialogue participants to mention those actions to specify a puzzle piece. In contrast, in this experimental setting, the evaluators are just overhearers who do not actually solve the puzzle, thus pay less attention to the actions in the video. This overhearer effect might explain the evaluators' failure in resolving AMEs.

5.3.2 Comparison of the TP and FN instances

While the results for the two types of the FP instances are very similar, there is a notable difference in the results of the TP and FN instances. The average IR of the TP instances is about 25 % higher than that of the FN instances. In terms of majority votes, the majority of evaluators was able to correctly identify the referent of all 20 TP instances, while this was only the case for 16 out of the 20 FN instances. IR and MIR show a similar tendency.

We implemented a pairwise *t*-test on the result of the TP and FN instances and found a significant difference at 0.01 % level ($t(33) = 16.24$). This difference suggests that it is more difficult for humans to identify the referent of a DP in the FN context than in the TP context. This difficulty in the FN context would be remedied by using definite noun phrases rather than DPs. Thus, our classifier's suggestion, the non-use of DPs in such FN contexts, should not necessarily be considered as wrong, if correct identification rather than human-likeness is concerned.

The difference in the agreement score for correctly answered instances also supports this claim. The overall agreement score of all TP instances is 0.969 ($SD = 0.049$), while that of the FN instances is 0.871 ($SD = 0.172$); there is a difference of about 10 %. The *SD* for the TP instances is smaller than that for the FN instances, reflecting the fact that the evaluators strongly agreed on the correct answers for the TP instances.

The majority of the evaluators successfully identified the correct referent for all TP instances. There were 19 wrong responses to the TP instances among a total of 680 responses. These 19 responses were distributed over 10 out of the 20 instances. We could find no particular tendency nor concentration of wrong responses on a specific instance. In contrast, there were 4 FN instances for which a majority of evaluators was unable to correctly identify the target. Two of them started with: "*sore to onazi katati/ôkisa* . . . (the same shape/size as that . . .)" where a DP is a reference object to refer to another object. Such complex expressions might enhance the overhearer effect.

5.3.3 Identification time and repetitions

Table 8 shows the average identification time (IT) of referents. The IT of FP-Orig tends to be shorter than that of FP-DP, indicating that the evaluators identified a referent more quickly when provided with non-DP referring expressions as used in the corpus. A pairwise *t*-test for all expressions, however, showed no significant

Table 8 Average identification time (IT) [msec]

	TP	FN	FP-DP	FP-Orig
Correct responses	4,038	4,614	5,408	4,984
All responses	4,102	5,759	6,100	5,514

Table 9 Number of repeatedly watched instances

	TP	FN	FP-DP	FP-Orig
Correct responses	18	40	37	14
All responses	23	74	47	26

difference of the IT between the FP-DP and FP-Orig instances. But we note that this comparison obscures the fact that the expressions of the FP-DP instances are generally shorter than the original expressions. In fact, the original expression is on average 2.72 ($SD = 2.04$) Japanese character longer than its replacement DP ("*sore* (that)", two Japanese characters). Hence, if excluding the reading time from IT, the difference of the actual identification time between the FP-DP and FP-Orig instances might likely be bigger than what we can see in Table 8.

As shown in Table 9, regarding the number of times the evaluators repeatedly watched a video, they behaved differently for the FP-DP and FP-Orig instances. Although the number of correct responses to the FP-DP and FP-Orig instances are almost the same as shown in Table 7, the number of repeatedly watched FP-DP instances is 2.6 times larger than that of the FP-Orig instances, suggesting some of the FP-DP instances are more "difficult" for evaluators, increasing the number of repeats. This fact is also reflected in the longer IT of the FP-DP instances shown in Table 8.

Regarding the TP and FN instances, we also note a difference in identification time, in the same direction as the identification accuracy. Namely, the IT of the TP instances was between about 500–1,400 ms less than that of the FN instances, suggesting that the FN instances are more difficult for humans. We implemented a *t*-test on this difference for all expressions which showed significance at a 0.01 % level ($t(33) = 7.79$).

Comparing the number of the repetitions of the TP and FN instances, the number of the repetitions of the FN instances is more than twice that of the TP instances. This indicates that even when only considering correct responses, the FN instances are more difficult than the TP instances.

5.4 Reviewing the results of the intrinsic evaluation

Considering the experimental result that replacing original referring expressions in the corpus with a DP "*sore* (that)" for the FP instances does not degrade the identification accuracy by the evaluators, we can argue that the FP instances should be considered "correct" instead of "wrong" as far as identification accuracy is concerned.

Based on this understanding, we recalculate the evaluation results shown in Table 4. In the task-performance evaluation, we tested 36 out of the 133 FP

instances and in 34 out of 36 instances, a DP succeeded to make the evaluators identify the correct referent. Only considering these 34 instances "correct" increases the precision from 0.664 to 0.718, and F-measure from 0.730 to 0.762. In this recalculation of identification accuracy, we assume that the rest of TP instances in the corpus remain "correct". Some of the TP instances, however, could not be correctly resolved by the evaluators in the experimental setting described in this paper. In the future, we will need to conduct a task-performance evaluation of all TP instances as well in order to better understand the degree to which we can generalise the results of our experiment.

6 Conclusions

This paper described an extrinsic task-performance evaluation of referring expressions in situated collaborative dialogues. In our previous work (Spanger et al. 2010b), we constructed a classifier which suggested the use of demonstrative pronouns (DPs) to refer to a given target in a given situation of a dialogue, where the intrinsic evaluation of the classifier based on corpus-matching was also conducted and discussed. A detailed analysis of this intrinsic evaluation results led to the question of whether the discrepancies between the classifier's output and the human-generated expression in the corpus might not necessarily mean the wrong decision of the classifier. This question motivated us to conduct the extrinsic task-performance evaluation reported in this paper.

In the task-performance evaluation, we particularly focused on the false positive (FP) instances where the classifier predicted using a DP but a dialogue participant had not used a DP in the corpus. Replacing the original expression with a DP in these FP instances, we presented different evaluator groups with the modified instances and the original ones in the same dialogue context. Comparing the evaluators' responses in terms of the identification rate (IR) and the majority identification rate (MIR), we found that these DPs perform as effectively as the original expressions for human evaluator to identify their referents. Based on the results, we reviewed our earlier evaluation results of the intrinsic evaluation and showed the F-measure of the classifier increased from 0.730 to 0.762. This result is in line with the discussion on intrinsic and extrinsic evaluations of referring expressions (Belz and Gatt 2008; Belz et al. 2010; Krahmer and van Deemter 2012) that claims each evaluation method measures two different aspects: a speaker-oriented aspect (human-likeness) and a hearer-oriented aspect (optimality for referent identification), and both should be taken into account for evaluating systems.

An interesting finding is that the evaluators could not perfectly identify the correct referents even when the original expressions used in the corpus were presented. We might be able to attribute these failures to the fact that the evaluators were external to the ongoing task, i.e. the overhearer effect. This result suggests a necessity to suppress the overhearer effect in the design of task-performance evaluations.

We adopted the identification time (IT) measures as well in order to consider the efficiency of expressions. We found that IT tends to be slightly longer for the DPs. This might be one of the reasons the dialogue participants in the corpus used more explicit referring expressions than DPs, which reduce the cognitive load of the hearers, hence induce a quicker response by the evaluator.

We also compared the true positive (TP) instances and the false negative (FN) instances. Note that DPs are used in the corpus in both types of instances, but the classifier could correctly predict DPs for only the TP instances. Our evaluation measures suggested that the FN instances were more difficult than the TP instances for the human evaluators to identify their referents. This provides indirect evidence that the classifier's suggestion of non-DPs in the FN instances is not necessarily wrong in terms of optimality for referent identification. Since our current classifier does not provide any concrete expression when suggesting non-DPs, a task for future work is to create a full model of generating referring expressions before conducting a full-fledged task-performance evaluation of referring expressions.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The HCRC map task corpus. *Language and Speech, 34*(4), 351–366.

Belz, A., & Gatt, A. (2008). Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-08: HLT, Short Papers* (pp. 197–200).

Belz, A., & Kow, E. (2010). The GREC challenges 2010: Overview and evaluation results. In *Proceedings of the 6th international natural language generation conference* (pp. 219–229).

Belz, A., Kow, E., Viethen, J., & Gatt, A. (2010). Referring expression generation in context: The GREC shared task evaluation challenges. In E. Krahmer, & M. Theune (Eds.), *Empirical methods in natural language generation* (Vol. LNCS5790, pp. 294–327). Berlin: Springer.

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

Byron, D., Koller, A., Striegnitz, K., Cassell, J., Dale, R., Moore, J., et al. (2009). Report on the first NLG challenge on generating instructions in virtual environments (GIVE). In *Proceedings of the 12th European workshop on natural language generation (ENLG 2009)* (pp. 165–173).

Cahill, A., & van Genabith, J. (2006). Robust PCFG-based generation using automatically acquired lfg approximations. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics* (pp. 1033–1040).

Campana, E., Tanenhaus, M. K., Allen, J. F., & Remington, R. (2011). Natural discourse reference generation reduces cognitive load in spoken systems. *Natural Language Engineering, 17*(3), 311–329.

Carenini, G., & Moore, J. D. (2006). Generating and evaluating evaluative arguments. *Artificial Intelligence, 170*(11), 925–952.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1–39.

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the 27th annual meeting of the association for computational linguistics* (pp. 68–75).

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

Di Eugenio, B., Glass, M., & Trolio, M. J. (2002). The DIAG experiments: Natural language generation for intelligent tutoring systems. In *Proceesings of the 2nd international natural language generation conference (INLG 2002)* (pp. 120–127).

Di Eugenio, B., Jordan, P. W., Thomason, R. H., & Moore, J. D. (2000). The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human-Computer Studies, 53*(6), 1017–1076.

Foster, M. E., Giuliani, M., & Knoll, A. (2009). Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of the 47th annual meeting of the ACL and the 4th IJCNLP of the AFNLP* (pp. 879–887).

Gargett, A., Garoufi, K., Koller, A., & Striegnitz, K. (2010). The GIVE-2 corpus of giving instructions in virtual environments. In *Proceedings of the seventh conference on international language resources and evaluation (LREC 2010)* (pp. 2401–2406).

Gatt, A., & Belz, A. (2010). Introducing shared tasks to NLG: The TUNA shared task evaluation challenges. In: E. Krahmer, & M. Theune (Eds.), *Empirical methods in natural language generation* (Vol. LNAI 5790, pp. 264–293). Berlin: Springer.

Gupta, S., & Stent, A. J. (2005). Automatic evaluation of referring expression generation using corpora. In *Proceedings of the 1st workshop on using Corpora in NLG*.

Heeman, P. A., & Hirst, G. (1995). Collaborating on referring expressions. *Computational Linguistics, 21*(3), 351–382.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 91–117.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods: Support vector learning* (pp. 169–184). Cambridge: MIT-Press.

Jordan, P. W., & Walker, M. A. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research, 24*, 157–194.

Khan, I., van Deemter, K., Ritchie, G., Gatt, A., & Cleland, A. A. (2009). A hearer-oriented evaluation of referring expression generation. In *Proceedings of the 12th European workshop on natural language generation (ENLG 2009)* (pp. 98–101).

Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., et al. (2010). Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). In *Proceedings of the 6th international natural language generation conference* (pp. 243–250).

Krahmer, E., & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics, 38*(1), 173–218.

Lester, J. C., Voerman, J. L., Towns, S. G., & Callaway, C. B. (1999). Deictic believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence, 13*(4–5), 383–414.

Mitkov, R. (2002). *Anaphora resolution*. London: Longman.

Ono, K. (1994). Territories of information and Japanese demonstratives. *The Journal of the Association of Teachers of Japanese, 28*(2), 131–155.

Papineni, K., Roukos, S., Ward, T., & Jing Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics (ACL 2002)* (pp. 311–318).

Paraboni, I., van Deemter, K., & Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational Linguistics, 33*(2), 229–254.

Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics, 35*(4), 529–558.

Reiter, E., Robertson, R., & Osman, L. M. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence, 144*(1–2), 41–58.

Reiter, E., & Sripada, S. (2002). Should corpora texts be gold standards for NLG? In *Proceedings of the 2nd international natural language generation conference (INLG 2002)* (pp. 97–104).

Reiter, E., Sripada, S., Hunter, J., Yu, J., & Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence, 167*(1–2),137–169.

Spanger, P., Iida, R., Tokunaga, T., Teri, A., & Kuriyama, N. (2010a). Towards an extrinsic evaluation of referring expressions in situated dialogs. In J. Kelleher, B. M. Namee, & I. van der Sluis (Eds.), *Proceedings of the sixth international natural language generation conference (INGL 2010)* (pp. 135–144).

Spanger, P., Yasuhara, M., Iida, R., & Tokunaga, T. (2009). Using extra linguistic information for generating demonstrative pronouns in a situated collaboration task. In *Proceedings of PreCogSci 2009: Production of referring expressions: Bridging the gap between computational and empirical approaches to reference*.

Spanger, P., Yasuhara, M., Iida, R., Tokunaga, T., Terai, A., & Kuriyama, N. (2010b). REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources and Evaluation, 46*(3), 461–491.

Sparck Jones, K., & Galliers, J. R. (1996). *Evaluating natural language processing systems: An analysis and review*. Berlin: Springer.

Stoia, L., Shockley, D. M., Byron, D. K., & Fosler-Lussier, E. (2006). Noun phrase generation for situated dialogs. In *Proceedings of the 4th international natural language generation conference (INLG 2006)* (pp. 81–88).

Striegnitz, K., Denis, A., Gargett, A., Garoufi, K., Koller, A., & Theune, M. (2011). Report on the second second challenge on generating instructions in virtual environments (GIVE-2.5). In *Proceedings of the 13th European workshop on natural language generation (ENLG 2011)* (pp. 270–297).

Tokunaga, T., Iida, R., Terai, A., & Kuriyama, N. (2012). 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)* (pp. 422–429).

van Deemter, K., Gatt, A., van der Sluis, I., & Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science, 36*(5), 799–836.

van der Sluis, I., Gatt, A., & van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. In *Proceedings of recent advances in natural languae processing (RANLP 2007)*.

van der Sluis, I., & Krahmer, E. (2007). Generating multimodal references. *Discourse Processes, 44*(3), 145–174.

Vapnik, V. N. (1998). *Statistical learning theory, adaptive and learning systems for signal processing communications, and control*. New York: Wiley.

Young, R. M. (1999). Using Grice's maxim of quantity to select the content of plan descriptions. *Artificial Intelligence, 115*, 215–256.