Dialogue Reference in a Visual Domain

Jette Viethen[1], Simon Zwarts[1], Robert Dale[1], Markus Guhe[2]

[1]Centre for Language Technology
Macquarie University
Sydney, Australia
[jviethen|szwarts|rdale]@science.mq.edu.au

[2]School of Informatics
University of Edinburgh
Edinburgh, Scotland
m.guhe@ed.ac.uk

Abstract

A central purpose of referring expressions is to distinguish intended referents from other entities that are in the context; but how is this context determined? This paper draws a distinction between discourse context –other entities that have been mentioned in the dialogue– and visual context –visually available objects near the intended referent. It explores how these two different aspects of context have an impact on **subsequent reference** in a dialogic situation where the speakers share both discourse and visual context. In addition we take into account the impact of the reference history –forms of reference used previously in the discourse– on forming what have been called **conceptual pacts**. By comparing the output of different parameter settings in our model to a data set of human-produced referring expressions, we determine that an approach to subsequent reference based on conceptual pacts provides a better explanation of our data than previously proposed algorithmic approaches which compute a new distinguishing description for the intended referent every time it is mentioned.

1. Introduction

There is now a quite substantial body of work in referring expression generation, a subfield of natural language generation (NLG) where the central concern is how to determine the semantic content of a nominal expression whose purpose is to identify an intended referent for a hearer. In almost all of this work, the basic idea pursued is that, to be successful, a referring expression must distinguish the intended referent from other entities in the **context**; we call such a referring expression a **distinguishing description**.

One question that has not been explored a great deal in the literature is the question of how the set of entities that make up the context is determined. Clearly, getting the context set right is a key factor in how well such approaches to reference work: if the context set mistakenly contains entities that are not considered to be part of the context by the hearer, there is the danger that the referring expression generated may contain unnecessary information, potentially leading to unintended Gricean implicatures, or at the very least, wasted effort on the part of the generator. On the other hand, if the context set used by the generator does not contain entities that the hearer believes to be in the context, then the referring expression generated may be ambiguous from the hearer's point of view.

In the initial characterisations of this problem, it was recognised that the context, particularly for subsequent references to an entity already mentioned, would be determined by properties of the preceding *discourse* (see, for example, Dale, 1989). A number of influential works in the 1980s (for example, Grosz et al., 1983; Grosz and Sidner, 1986) were concerned with the hierarchical structure of discourse, and how this hierarchy impacted on the accessibility of previously mentioned entities at any given point in the discourse, thus determining which entities would serve as potential distractors for an intended referent. We can think of a context set derived in this way as consisting of what we will call **discourse distractors**.

Keeping track of discourse distractors is important for any natural language generation system whose purpose is to produce multisentential or more extended discourse, as in report generation or summarisation. Most recent work in referring expresion generation, however, has concerned itself with corpora of human-produced referring expressions elicited in situations where the context, and therefore the set of distractors, is defined in terms of a *visual* field. In the TUNA corpus (van der Sluis et al., 2006), subjects were asked to distinguish a target referent from the rest of the entities visible on the screen serving as distractors; similarly, in the GRE3D3 corpus (Viethen and Dale, 2008), subjects were asked to distinguish a target entity from two other entities visible in a simple blocksworld scene. In these cases, the context set consists of what we will call **visual distractors**.

An important and relevant orthogonal distinction which has not been considered as a factor in more recent work in this area is the distinction between **initial reference** and **subsequent reference**: we use an initial reference to introduce an entity into a discourse, and subsequent reference to refer to it after it has been introduced. This distinction is of considerable significance in the generation of extended discourse: for example, notwithstanding rare cases of cataphora, pronouns can be used for subsequent reference but not initial reference. The initial vs subsequent reference distinction is not of obvious significance in the visual reference studies mentioned above, however, since in these cases every reference is an initial reference: in the acquisition of the data, each scene is presented to subjects anew, and only one reference is generated per scene.

This raises the question of how subsequent reference works in a visual domain, where the visible objects are referred to repeatedly. One simple hypothesis, for example, would be that the initial reference is determined on the basis of visual distractors, and then subsequent references are determined on the basis of discourse distractors.

However, work in psycholinguistics suggests a rather different perspective. The literature in this area has explored

how **alignment**, whereby a conversational participant will adopt the same semantic, syntactic and lexical alternatives as the other party in a dialogue, has an impact on reference, with speakers forming **conceptual pacts** in their use of language (Brennan and Clark, 1996). The implication of much of this work is that one speaker introduces an entity by means of some description, and then (perhaps after some negotiation) both conversational participants share this form of reference, or a form of reference derived from it, when they subsequently refer to that entity.

In our work, we aim to model a conversational participant in a dialogue that concerns a shared visual scene, a common scenario where automated referring expression generation might be particularly useful. Consequently, we need to determine how these three factors—discourse context, visual context and the other participant's use of language—interact to determine how our model should construct appropriate referring expressions. In this paper we describe some initial experiments that explore this question.

Consideration of situations where all three factors play an explicit role is rare. Jordan and Walker (2000; 2005) explored how the model proposed by Grosz and her colleagues plays together with the task and the intentions of the speaker in influencing the content of non-pronominal subsequent referring expressions. However, unlike Jordan and Walker's domain, in which the objects under discussion were not usually visible to both dialogue partners, the landmarks on the maps that are described in our corpus are in general visible to both partners. A second difference is that at least the vast majority of the referring expressions in our corpus have the main function of uniquely identifying the target referent, which is only one of a number of tasks the descriptions in Jordan and Walker's corpus perform.

The remainder of this paper is organised as follows. The data set we use and the particular referring expressions we concentrate on are described in Section 2. Our aim is to explore the influence of different factors on the content of a subset of the subsequent referring expressions in our data with the ultimate goal of creating a system that is able to regenerate these expressions as accurately as possible. The hypotheses we set out to test are detailed in Section 3 and our model is outlined in Section 4.

In Section 5 we try out a variety of parameter settings for the model and compare the results to the human-produced instances in our data set using the DICE metric (Salton and McGill, 1983) and the MASI metric (Passonneau, 2006), which have been used in the recent REG Challenges (Gatt et al., 2008). We focus in particular on the first subsequent reference by the second speaker in the dialogue; our results show that the most influential factor shaping the content of the subsequent referring expressions from our data set is the way the target referent has been described previously in the dialogue, consistent with an alignment-based model. Combining this with the need to distinguish from either the set of other entities described in the discourse or the set of visual distractors does not considerably improve the results of our model, suggesting that the model assumed in earlier work on referring expression generation, where each reference is explicitly constructed with a deliberate concern for its ability to discriminate, is not the most explanatory.

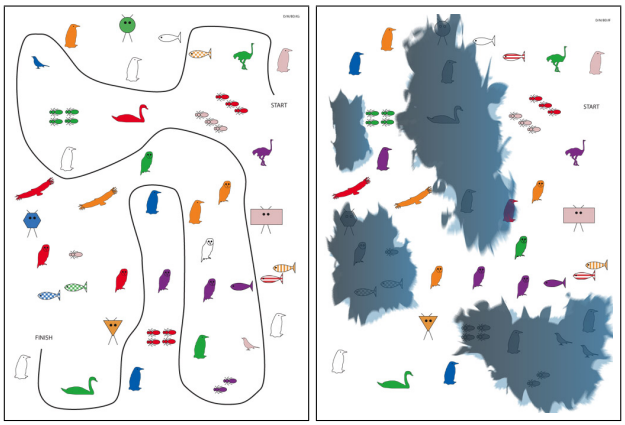

Figure 1: An example pair of maps.

We close with a summary of our conclusions and a number of suggestions for future work in Section 6.

2. The Data

We use the iMAP corpus (Louwerse et al., 2007; Guhe and Bard, 2008), which consists of 256 dialogues elicited from 32 dyads of participants. In each dialogue, one participant is the instruction giver (IG) and the other the instruction follower (IF). Both participants have a map with landmarks in front of them. The task is for the IG to communicate a path, which is only shown on her map, to the IF in such a way that the IF can draw it onto his map. This is complicated by the presence of a number of small discrepancies between the IG's and the IF's maps in the form of ink damage obscuring colour in some spots on the IF's map, and by slight differences in the landmarks. Figure 1 shows an example of a pair of maps.

The landmarks on the maps can be distinguished from each other by their COLOUR and one of the properties SHAPE, PATTERN, NUMBER or KIND, depending on the TYPE of landmark, which is also often included in the descriptions. Each map has one main type of landmark, which is either BUILDING, BIRD, TREES, BUGS, ALIEN, TRAFFIC SIGN, FISH or CAR. The main type of landmark on the maps in Figure 1 is BIRD; in the iMAP domain, birds can be distinguished by their COLOUR and their KIND.

We would like to develop a model of how referring expressions are constructed in this context. In this paper, we focus on the specific case of the first reference made by the second speaker to an entity that has already been introduced in the discourse by the other speaker. We call these descriptions **Second-speaker Initial References** (SSIRs) and the landmark they refer to in each case the **target landmark**. In this way, we aim to determine the influence of the three factors introduced above: the discourse context, the visual context, and the other speaker's particular use of language in describing the entity.

We have extracted a data set of 2579 instances of SSIRs from the corpus, against which we evaluate our model. Our focus here is on singular reference, although the corpus also

contains another 2366 coreference chains in which the target landmark was described at least once as part of a plural referring expression describing a group of several objects, such as *the two purple owls*.

3. Possible Scenarios

Given the three factors we have identified, there are three baseline hypotheses we can consider:

1. It could be the case that the only important factor is the **reference history** of the target referent—given that we are focussing on the first reference by the second speaker, this is the collection of descriptions that the first speaker to mention the target landmark used to describe it.

2. It might be the case that the reference history is not taken into account at all, but rather the SSIR only distinguishes from other entities that have recently been mentioned, the **discourse distractors**.

3. SSIRs might be similar to the first mentions by the first speaker in that they distinguish the target landmark from the **visual distractors** around it.

It is also possible that combinations of any two of these factors might be at play:

4. An SSIR might distinguish the target landmark from both discourse and visual distractors, but not take any account of the reference history.

5. While taking into account the reference history, an SSIR might also need to distinguish the target landmark from the visual distractors.

6. An SSIR might be based on the previous reference history, but it might also distinguish the landmark from discourse distractors.

And finally, all three factors might have an effect:

7. An SSIR might build on the content of the previous mentions in the reference history and also distinguish from both discourse distractors and visual distractors.

Intuitively, one might expect Hypothesis 6 to be the most plausible of these, on the basis that the reference history 'costs in' the discourse and visual distractors up to that point, so that a subsequent reference only needs to take additional account of any distractors mentioned since the target landmark was last mentioned.

4. The Model

We use a simple rule-based model to explore the scenarios from the previous section with the aim of creating a system that is able to re-generate the descriptions in our human-produced data set as accurately as possible. We define a number of input parameters that allow us to fine-tune our model and explore if and explicitly in which ways the factors mentioned above influence the content of the SSIRs we attempt to replicate.

Two parameters define if and how the reference history impacts an SSIR:

- `history_strategy`: This parameter toggles between a setting in which the SSIR repeats the last mention from the reference history in its entirety (`repeat_last`), a setting which retains only the head noun from that last mention (`drop_to_type`), and a setting which ignores the reference history (`drop_all`).

- `property_strategy` defines which properties can be used in the SSIR for the target landmark. They are either drawn from the set of previously `mentioned` properties, from the set of properties that the database holds for the target landmark (`db`; this may include properties which are visually available but have not yet been verbalised), or from both these sets (`union`).

One parameter controls the influence of the discourse distractors:

- `discourse_strategy` defines which properties of the discourse distractors the SSIR needs to distinguish from. We can take into account only the properties that have been `mentioned` for these landmarks, we can also take into account any additional visual properties that our database holds (`union`), or we can ignore the discourse distractors (`none`).

We define one parameter for the control of the visual distractors:

- `visual_strategy` defines whether the visual distractors get taken into account or not. We define a radius around the landmarks in such a way that each circle contains on average six visual distractors.[1]

The main algorithm used by our model is shown in Figure 2. Step #2 here ensures that the description rules out all distractors by applying the Incremental Algorithm (Dale and Reiter, 1995) (IA). We also experimented with the Greedy Algorithm (Dale, 1989), but in line with earlier experiments on initial reference (Viethen and Dale, 2006; van der Sluis et al., 2007), the IA selection strategy consistently outperformed the Greedy strategy.

In a post-processing step, we ensure that the TYPE of the target landmark is included in the SSIR, if it has not already been chosen by the model; this is a standard procedure in the literature to ensure that the resulting noun phrase will have a head. If no properties have been chosen, the TYPE is specified as PRON to ensure that the expression will be pronominalised.

5. Exploration

5.1. Method

We report performance in terms of accuracy and two set distance measures. Accuracy measures the proportion of times

[1] This is somewhat arbitrary; however, there is no straightforward way to determine which entities would be in the visual field of attention at any given point in time. We believe six is a reasonable average number based on our analysis of the maps.

	history strategy	property strategy	discourse strategy	visual strategy	average DICE	summed DICE	average MASI	Accuracy	strict subset	strict superset
1	repeat_last	n/a	none	no	0.673	0.680	0.531	40.2%	13.8%	33.1%
2	drop_to_type	n/a	none	no	0.459	0.457	0.339	22.3%	56.5%	6.9%
3	drop_all	db	mentioned	no	0.467	0.471	0.326	22.8%	30.9%	9.6%
4	drop_all	db	none	yes	0.558	0.570	0.348	18.6%	10.7%	39.1%
5	repeat_last	mentioned	none	yes	0.667	0.675	0.519	38.4%	12.8%	35.8%
6	repeat_last	union	none	yes	0.672	0.680	0.530	40.1%	13.7%	33.4%
7	drop_all	db	mentioned	yes	0.559	0.573	0.349	18.9%	9.1%	39.8%
8	drop_all	db	union	no	0.564	0.579	0.349	18.2%	8.6%	42.7%
9	repeat_last	union	mentioned	no	0.676	0.682	0.529	39.6%	12.8%	33.8%
10	repeat_last	mentioned	mentioned	no	0.677	0.681	0.528	39.4%	12.2%	34.94%
11	repeat_last	union	union	no	0.677	0.683	0.530	39.7%	12.6%	34.0%
12	repeat_last	mentioned	union	no	0.674	0.680	0.525	39.0%	11.9%	35.5%
13	repeat_last	union	mentioned	yes	0.675	0.682	0.528	39.4%	12.5%	34.4%
14	repeat_last	mentioned	mentioned	yes	0.671	0.677	0.518	37.9%	11.4%	36.9%
15	repeat_last	union	union	yes	0.676	0.682	0.529	39.5%	12.5%	34.4%
16	repeat_last	mentioned	union	yes	0.670	0.677	0.516	37.7%	11.4%	37.1%

Table 1: **The results for the parameter settings we explored.** Lines 1 to 4 are the baselines which only take into account one of the three factors. Lines 5 to 12 combine two factors each and Lines 13 to 16 combine all three factors.

```
    # Initialise
1   P = {the set of properties of the target landmark
    according to property_strategy}
2   C = {the set of distractors according to
    discourse_strategy and visual_strategy}
3   SSIR = {}

    # Step 1 — continue the history
4   SSIR = {the properties returned by the
    history_strategy}
5   C = C − {the distractors ruled out by SSIR}

    # Step 2 — rule out distractors
6   run_IA(SSIR,P,C)
7   return SSIR
```

Figure 2: The main algorithm

the system output is an exact match of the human-produced referring expressions. Both the DICE coefficient and the MASI scores are set-comparison metrics that deliver values ranging between 0 and 1. In the context of content determination in REG, they are applied by comparing the set of properties contained in the description that the system has produced to those contained in the human-produced reference description. The two measures have both been used to evaluate the fit of REG output to human-produced data in the recent evaluation challenges in this field (Gatt et al., 2008). The main difference between them is that MASI is biased in favour of solutions that are a subset or a superset of the gold standard.

Given two sets of attributes, A and B, DICE is computed as

$$\text{DICE} = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (1)$$

and MASI as

$$\text{MASI} = \delta \times \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

where δ is a monotonicity coefficient which biases the metric in favour of solutions that are a subset or a superset of the gold standard. It is defined as

$$\delta = \begin{cases} 0 & \text{if } A \cap B = \emptyset \\ 1 & \text{if } A = B \\ \frac{2}{3} & \text{if } A \subset B \text{ or } B \subset A \\ \frac{1}{3} & \text{otherwise} \end{cases} \quad (3)$$

We apply the metrics on the level of attributes: a property is counted as correct if it was included by our system and the gold standard SSIR also contained this property; the value of the property is not taken into account. In doing this we follow the practice of the TUNA evaluation challenges (Gatt et al., 2008), in order not to penalise our system's output in cases where the participants used property values that are superficially different from those in the data base and that our system therefore has no access to.

For both measures we report the average value over the complete data set. For DICE we also report what we call the **summed mean**, which is sensitive to the difference between getting a long referring expression right and getting a short one right. Like the BLEU metric used in MT (Papineni et al., 2002), it sums the numerators and the divisors separately. This ensures that a long referring expression contributes more to the overall score than a short one.

5.2. Results

The first four columns of Table 1 describe the parameter settings we explored; the columns in the centre show the DICE, MASI and Accuracy scores that each setting achieved; and the last three columns show information that is helpful for an error analysis.

Lines 1 to 4 show the results of the parameter settings that we used to test the baseline hypotheses from Section 3 (reference history only; discourse distractors only; and visual distractors only—recall that the database contains only those properties which are visually available). Lines 1 and 2 represent different ways in which we can take the reference history into account without giving the discourse or visual distractors any influence. They differ in the setting for `history_strategy`.[2] Setting this to `repeat_last`, i.e. simply repeating the last description used by the first speaker, achieves by far the best results here.

Line 3 represents the hypothesis that only the mentioned properties of the discourse distractors are taken into account in an SSIR. It does not reach the same DICE scores as those of Setting 1, but achieves similar results as Setting 2, which simply uses only TYPE.

Line 4 shows that only distinguishing from the visual distractors of the target landmark also results in worse performance than simply repeating the last mention. However, this does increase performance over only taking into account discourse distractors (Setting 3).

The performance increases again to the same level as that of Setting 1, if we combine taking into account visual distractors (`visual_strategy=yes`) with the best performing setting for the `history_strategy` (Lines 5 and 6). Making available both mentioned and visual properties achieves slightly better DICE scores than only using the properties of the target landmark that were mentioned in the history, but results in a small drop in Accuracy.

Combining the two types of distractors and not giving the reference history any influence (Lines 7 and 8) achieves worse results than any of the settings that are based on repeating the last mentioned properties.

The settings in Lines 9 to 12 combine the influence of the reference history with that of the discourse distractors. For both the target landmark and the discourse distractors we test using only mentioned attributes and additionally using all properties that are recorded in the data base. There is almost no difference between these four settings. The best performance is achieved here by the settings in Line 11, which also outperform the four best settings in which we tried combining all three factors (Lines 13 to 16).

Overall the summed DICE scores are slightly better than the averaged DICE results. This shows that we are getting some of the longer referring expressions right, which the average does not honour as much as the summed DICE. The MASI scores cannot be directly compared to the DICE scores, although we note that the ranking of the settings according to MASI is only very slightly different from the DICE ranking. However, MASI does correspond better to the Accuracy ranking than DICE does.

The principal finding here is that the content of the previous mentions of the target landmark have a large impact on the content of an SSIR. Taking into account the visual distractors does not improve the performance of our model; and, perhaps surprisingly, the need to distinguish from distractors introduced in the discourse also does not seem to have a large influence on the content of the SSIRs we attempt to replicate.

[2]Changing `property_strategy` would have no effect as no distractors have to be excluded in these settings.

5.3. Error Analysis

The last three columns of Table 1 show how often each setting produced exactly the same referring expression as in the human-produced data set, and how often our result was either a *strict subset* or a *strict superset* of the human-produced referring expression. When our model produces a referring expression whose content is a strict subset of that of the human-produced description, we may be guilty of having produced an underspecified description given the context; and when we produce a superset, our referring expression will be overspecified.

These columns show that whenever our model achieves relatively high results, both for Accuracy and DICE score, it is also overspecifying to a large degree. This means that the remaining error is largely due to the system including too many properties in the referring expressions it generates.

This is in particular true for the simple baseline that copies the last mention of the target landmark. What we can conclude from these results is that the 14% of human-produced SSIRs that are a superset of the last-used description do not contain this additional information in order to distinguish the target landmark from any discourse or visual distractors. If this was the case, we should see a real increase in performance from the baseline setting under which our system simply copies the last mention (Line 1 in Table 1) to those settings where it takes the last mention as a basis but adds extra properties to it if that is necessary to distinguish from new distractors (Lines 5, 6 and 9 to 12).

6. Conclusions and Future Work

Our experiments indicate that, for second speaker initial reference at least, the reference history has a large impact, while visual distractors do not play an important role. Perhaps surprisingly, taking into account discourse distractors does not appear to help much either. The simple baseline of copying the last mention gets close to the best results.

This suggests a general strategy for a dialogue agent might be to simply copy the form of reference used by their conversational participant; as already noted, this observation is a confirmation of the idea that underlies research on alignment, and is consistent with findings in the psycholinguistic literature. In the context of computational work on referring expression generation, however, this finding is significant because it stands in opposition to the view that has been tacitly accepted for close to two decades, that an algorithm is required to compute the content of a subsequent referring expression. Of course, in hindsight such a finding is not so surprising: as noted earlier, previous references to the entity already 'cost in' existing distractors, whether they arise from the discourse or the visual context, and so in effect the other speaker has already done all the work. When an agent needs to refer to some entity already introduced by the other party, a change in referential form only needs to be considered if the context of reference has changed in

some relevant way. Each reference to a given entity in a discourse can be thought of as 'building on previous work' in this way, opening the door to more collaborative algorithms for referring expression generation.

Our experiments so far simplify various aspects of the scenarios we are exploring here. First, we could take into account more situational factors; in particular, in the iMAP corpus, we should take account of whether the target's colour is obscured by an inkblot on the IF's map, as well as features such as the particular speaker or speaker-pair, and the main type of landmarks on the map. While these features might increase performance with regard to this particular data set, it would be at the cost of a loss of generality of the model. However, taking into account the speaker is a necessity if we take seriously the view that speaker variation is a significant factor underlying variation in the content of referring expressions (Viethen and Dale, 2006).

Also, additional discourse factors could be integrated. For example, explicit communication between the dialogue partners about the areas in which colour is missing for the IF, or which of the landmarks seem to differ between the maps, could be taken into account. Another discourse factor that might have an impact on referring expression generation is the dialogue act type of each utterance containing a referring expression; however, exploring these aspects would require considerable additional annotation of the corpus.

Currently the decision as to whether our model takes account of the reference history is somewhat arbitrary. Instead, it might make sense to make the reference history's influence dependent on the distance between the current referring expression and the last mention of the target referent. It is likely that the impact of this last mention diminishes the further back it lies. It is also likely that a larger number of discourse distractors being mentioned since the last reference to the target landmark decreases its impact.

The SSIRs, which we have focussed on in this work, are, of course, only one subclass of coreferential expressions in dialogues. As well as improving the performance of our model on this special case, we aim to explore how the model extends to other subsequent references in the corpus.

7. References

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

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

Robert Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver B.C., Canada.

Albert Gatt, Anja Belz, and Eric Kow. 2008. The TUNA Challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Conference on Natural Language Generation*, pages 198–206, Salt Fork OH, USA.

Barbara J. Grosz and Candance L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 44–49, MIT, Cambridge MA, USA, 15-17 June.

Markus Guhe and Ellen Gurman Bard. 2008. Adapting referring expressions to the task environment. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 2404–2409, Austin, TX. Cognitive Science Society.

Pamela W. Jordan and Marilyn Walker. 2000. Learning attribute selections for non-pronominal expressions. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Hong Kong, China.

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

Max M. Louwerse, Nick Benesh, Mohammed E. Hoque, Patrick Jeuniaux, Gwyneth Lewis, Jie Wu, and Megan Zirnstein. 2007. Multimodal communication in face-to-face computer-mediated conversations. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 1235–1240.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia PA, USA.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York NY, USA.

Ielka van der Sluis, Albert Gatt, and Kees van Deemter. 2006. Manual for the TUNA corpus: Referring expressions in two domains. Technical Report AUCS/TR0705, Computing Department, University of Aberdeen, UK.

Ielka van der Sluis, Albert Gatt, and Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

Jette Viethen and Robert Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation*, pages 63–70, Sydney, Australia, July.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the 5th International Conference on Natural Language Generation*, pages 59–67, Salt Fork OH, USA.