Adapting referring expressions to the task environment

Adapting Referring Expressions to the Task Environment

Markus Guhe (m.guhe@ed.ac.uk)
Department of Linguistics and English Language & Human Communication Research Centre,
40 George Square, Edinburgh, EH8 9LL, UK

Ellen Gurman Bard (ellen@ling.ed.ac.uk)
Department of Linguistics and English Language & Human Communication Research Centre,
40 George Square, Edinburgh, EH8 9LL, UK

Abstract

When people refer to objects linguistically, they must choose properties of the object that make it possible for the listener to identify the intended referent. We show that this selection of properties not only depends on the task environment but also changes over the course of time. We find that the salient feature *color* is used less often over time because of its limited utility in our task, while other features with high utility are used more often. We also find that the speaker does not change his/her behavior because of feedback from the interlocutor but because of experience gained when the roles in the task are reversed.

Keywords: referring expression; dialogue; demands of the task environment; adaptation

Introduction

A linguistic problem that has been receiving much attention in recent years is the use and generation of referring expressions. Referring expressions are linguistic expressions (usually a combination of nominal and prepositional phrases) that identify a referent entity the in real world (the target object) or a discourse object. The ostensible purpose of referring expressions is to distinguish the target from the concurrent distractor set. For example, in the set of objects in Figure 1, *the black cup* and *the small, black cup* would both be identify the cup at the lower left.

Figure 1: A distractor set

While computational approaches often try to generate expressions that uniquely and minimally select the target object, such algorithms are computationally costly. Furthermore, people (1) produce non-minimal expressions, which contain redundant information (e.g., Pechmann 1989) and (2) interpret such expressions more easily (e.g., Paraboni, van Deemter and Masthoff 2007).

A prominent account of how such non-minimal referring expression can be generated is the algorithm by Dale and Reiter (1995) for which many extensions have been developed (see van der Sluis (2005) for a recent overview). This algorithm incrementally tests whether using an attribute rules out distractor objects and in a fixed order. For the domain used in Figure 1, for example, this preference list could be <type, color, size>. Identifying the object to the right would then produce the non-minimal expression *large, white, cup* by first adding the type attribute (which has a special status and is always added), then by adding *white* (because it removes the object in the lower left from the distractor set), finally by adding *large* (because it removes the object in the top left). Once an attribute has been selected it will not be deselected again.

While these approaches deal with uptake of descriptive opportunities, they do not account for the adaptations that a speaker makes over time to the demands of the current task and the task environment. (For the purposes of this paper, we will use *task* to include both goal and environment.) Purposely, such paradigms lack history: A participant is usually presented with a picture like Figure 1 and instructed to produce a suitably distinguishing expression. The trial terminates with no feedback and is followed by others, which use different objects and distinguishing features. Here, we look at referring expressions in an unrestricted, task-oriented dialogue in which the interlocutors get natural feedback on failures of reference. We use a variant of the HCRC Map Task (Anderson et al. 1991) in which an Instruction Giver who can see the route on a schematic map communicates it to an Instruction Follower who must reproduce it. Several different features distinguish cartoon landmarks. We find that the use of those features changes over time. Elaborating on the results in Guhe and Bard (2007) we ask here how and why the use of feature terms in referring expressions changes over time.

Of particular interest is the change in the use of color terms, because color is a perceptually salient property and is usually one of the first few features tested in the incremental Dale and Reiter algorithms. In the present experiment, however, color is an unreliable distinguisher. But the participants do have another feature in each map that will produce adequate distinctions.

In this paper, we address the following questions:

1. Utility: What are the influences of utility on the choice of features for introductory referring expressions?

2. Audience design: Can an Instruction Giver adjust the use of color terms towards a level justified by the listener's problems, or is firsthand experience as an Instruction Follower what drives the change in behavior?
3. Specificity: Are features retained only in maps in which they are distinguishing, or does their local utility encourage their global use?

Comparison to existing research

The problem of whether the use of features changes with the demands of the task environment has scarcely been addressed in the literature.

Brennan and Clark's (1996) conceptual pacts are only partially applicable here, because they give an account of how speakers refer to objects after they have been introduced. Our questions here, however, address the choice of features for initial mentions of many different entities.

Though Garrod and Doherty (1994) show how a community of speakers establishes a sub-language for referring to complex entities, we are concerned with the internal structure of the referring expressions themselves when the entities are simple and independent. Ultimately we propose a utility-based explanation instead of one based on precedence and salience.

There is already evidence that extra-linguistic factors play a role in generating referring expressions. For example, Arnold and Griffin (2007) show that the presence of a second character influences the choice of whether to use a pronoun or the character's name for reference. This is true even if the characters have a different gender, so that the name does not disambiguate better than the pronoun. Arnold and Griffin attribute the effect to the cognitive load imposed by generating the referring expression. Their findings suggest that cooperative factors in dialogue (e.g. Clark 1996) compete with a more speaker-oriented forces (e.g. Bard et al. 2000). In this view, the speaker makes the general assumption that what he/she knows is also shared knowledge. Only if problems arise in the dialogue, e.g. by explicit feedback from the listener, might the speaker adapt to the listener's needs. The forces work in the same direction when an Instruction Follower in a communicative task becomes an Instruction Giver in a later trial (Haywood 2005). At this point, the speaker can tailor her presentation to a listener whose role she understands from first-hand experience.

Ideally, in fact, audience design in referring expressions ought to promote least collaborative effort (Clark & Wilkes-Gibbs 1986). Appropriately, overspecified referring expressions (Dale & Reiter 1995; Paraboni, van Deemter and Masthoff 2007; Pechmann 1989) both help the listener to identify the target object and permit the speaker to employ a generation process of greatly reduced complexity. Since both interlocutors benefit from using such referring expressions, the communicative strategy cannot be attributed uniquely to concern for the listener's needs.

Our task, however, was designed to make the salient feature, color, counterproductive in the many cases where it does not match between the players' maps. Though color is still required to make a distinction in some cases, there is an obvious cost attached which should grow over time. Thus the work contrasts with studies which use machine learning techniques to let global properties of linguistic corpora determine how attributes are selected for modified versions of the Dale and Reiter algorithm (Jordan and Walker 2005). Although these algorithms incorporate psychological findings (e.g., Brennan and Clark's conceptual pact model), they provide a starting point, not a continuously adaptable system.

Experiment

The iMAP experiment is a modified Map Task (Anderson et al 1991). The Map Task is an unscripted route-communication task in which an Instruction Giver and an Instruction Follower each have a map of the same fictional location. They collaborate to reproduce on the Follower's map a route shown only on the Giver's

Materials

Some landmarks differ between the two maps. In our experiment they can differ by:

1. Being absent on one of the maps or present on both;
2. Mismatching in a feature between the two maps (most notably color);
3. Being affected or not by 'ink damage' that obscures the color of some landmarks on the Instruction Follower's map.

There are four landmark features, each with two levels to distinguish:

1. Number (bugs, trees),
2. Pattern (fish, cars),
3. Kind (birds, houses/buildings),
4. Shape: (aliens, traffic signs).

There are three experimental variables:

1. Homogeneity: whether the landmarks on a map are of just one kind or whether the landmarks are of different kinds. The top two maps of Figure 2 differ in homogeneity.
2. Orderliness: whether the ink blot on the Instruction Follower's map obscures a contiguous stretch of the route (orderly) or a non-contiguous stretch (disorderly). The bottom two maps of Figure 2 differ in orderliness.
3. Animacy: whether the landmarks on a map are animate or inanimate (thus, on the mixed maps there are only landmarks from the 4 inanimate or the 4 animate kinds of landmarks).

Procedure

The participants are told that the maps are 'of the same location but drawn by different explorers'; they but are not told how or where the maps differ. They are instructed to recreate the route on the Instruction Follower's map as accurately as possible.

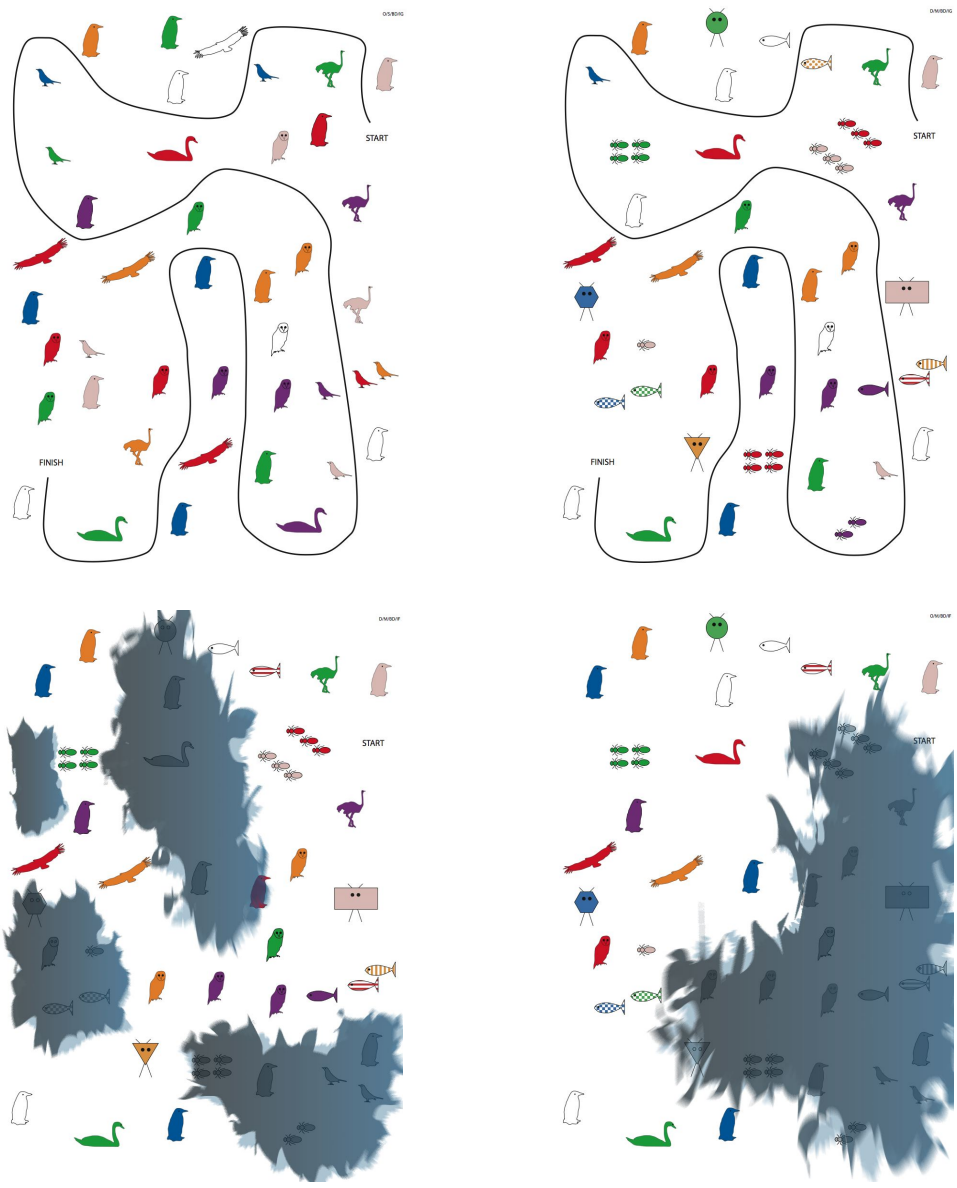

Figure 2: Example maps

Each dyad had to complete two simple training maps and then eight maps, one of each kind of landmarks. The maps were counterbalanced with respect to the experimental conditions. After the fourth map, the role of Giver and Follower exchanged roles.

To reduce the variability of referring expressions each participant had to name a few landmarks that would occur on the following map and was prompted textually how to name them. Landmarks were not labeled (see Figure 2).

Setup and data collection

Participants sat in front of individual computers, facing each other, but separated by a visual barrier (see Figure 3).

The communication was recorded using 5 camcorders. Eye gaze was recorded for the Information Giver only using a remote eye tracker. Speech was recorded using a Marantz PMD670 recorder whereby Instruction Giver and Instruction Follower were recorded on two separate (left and right) channels using two AKG C420 headset microphones. The speech was later transcribed manually. The routes drawn by the Instruction Followers were recorded by his/her computer.

As they were in the same room, participants could hear each other's speech, and they could see each other in the left half of their monitor, which showed the dialogue partner's upper torso video stream. The right half of the monitor showed the map (see Figure 4).

Figure 3: Picture of the experimental setup

Figure 4: Picture of the Instruction Follower's screen

Participants and data coding

Sixty-four undergraduates of the University of Memphis participated for course credits.

For the current analysis, the recorded dialogues were transcribed verbatim and all referring expressions were coded for use of color terms and for terms describing the landmark features (number, pattern, kind, shape).

Results

We analyzed only the. the referring expressions that mentioned a landmark for the first time in a dialogue. This restriction should reveal adaptation to the task environment without effects of conceptual pacts or of reduced referring expressions across repeated mentions. We also used only the introductions made by the Instruction Giver, who introduced landmarks in the majority of cases (7995 of 9567 landmarks or 83%). The analyses for the whole data set (introductions by Giver and Follower) give very similar results.

We split each dialogue into quartiles to see how the use of feature terms changed over the course of each dialogue, and specified each dialogue with respect to whether the Instruction Giver already had experience as Instruction Follower. Adaptation to current task demands would be reflected in a reduction in rate of the often unhelpful color terms over time, and either a constant or an increased use of other features.

For color terms, we performed a 3-way repeated measures ANOVA (experience(2) x map encountered(4) x quartile(4)) on the arcsine transformed proportion of introductory mentions containing color terms. For the other distinguishing features (number, pattern, kind, shape) we also performed a 3-way repeated measures ANOVA to check for specificity of change (experience(2) x quartile(4) x map(4)) again on the arcsine transformed proportion introductory mentions by the Instruction Giver that use the feature. We use the arcsine transform for the ANOVA because of potential problems ANOVAs have with non-normal distributions of mean values, cf. Jaeger (2007).

Feature relevance compared the use of each feature in maps where it expressed critical distinctions between landmarks with the rate in one of the map pairs where it was not relevant.

Changes across dialogues

Color As can be seen in Figure 5, after an initial drop after the first quartile of the first map, there is a sharp drop when the roles of the speakers change after the fourth map. Speakers use significantly fewer color terms when they already have experience as Instruction Follower than they do if they have been only Instruction Givers (0.267 color terms per landmark introduction in the first four maps vs. 0.175 in the second four maps; $F_1(1, 28) = 7.90$, $p < 0.01$).

Within a player's tenure as Instruction Giver there is no significant effect of maps, i.e., no gradual change from dialogue to dialogue.

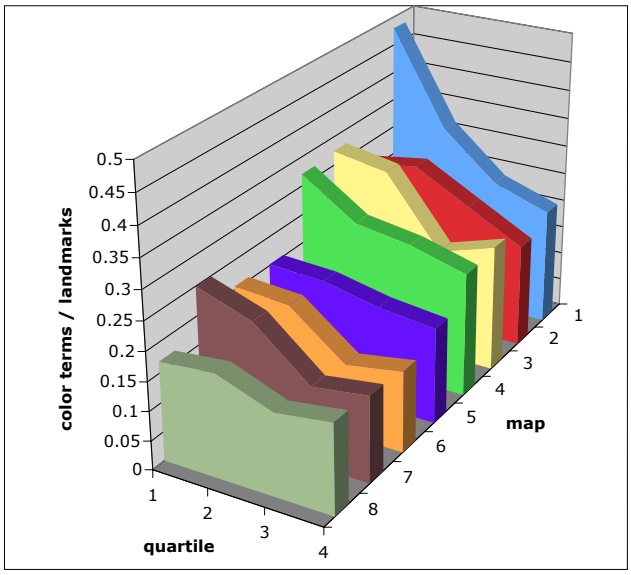

Figure 5: Change of the use of color terms over the course of the 8 maps and within maps

Useful features There is no significant effect of experience for the other feature terms on the maps where they distinguish landmarks ($F_1(1, 31) < 1$). Because the useful features are mentioned on average for 81.3% of the landmarks in their maps, this most likely is a ceiling effect. (Remember that on the mixed maps there are also landmarks for which the 'map feature' is not distinguishing, e.g. bugs on a bird map cannot be distinguished by kind but are still distinguished by number.)

Changes within dialogues

Color The use of color terms in introductory mentions significantly decreases within dialogues ($F_1(2, 54.8) = 15.57$, $p < 0.001$ means of the four quartiles (original proportions): q(1) = 0.280; q(2) = 0.252; q(3) = 0.190; q(4) = 0.188), cf. also Figure 7.

There is a significant interaction between experience and quartile ($F_1(2.3, 64.3) = 3.57$, $p < 0.05$), cf. Figure 6. The effect is due to (1) a steady decrease of number of color terms mentioned during the first four maps, i.e. before the roles are exchanged (significant posthoc differences between quartile-pairs 1–3, 1–4 and 2–4) and (2) significant differences of experience for quartiles 1, 2 and 3.

Clearly, the main within-map adaptation takes place during the first four maps. Givers with Follower experience (shown in red in Figure 6) initially use color terms at roughly the rate their interlocutors had employed late in map dialogues..

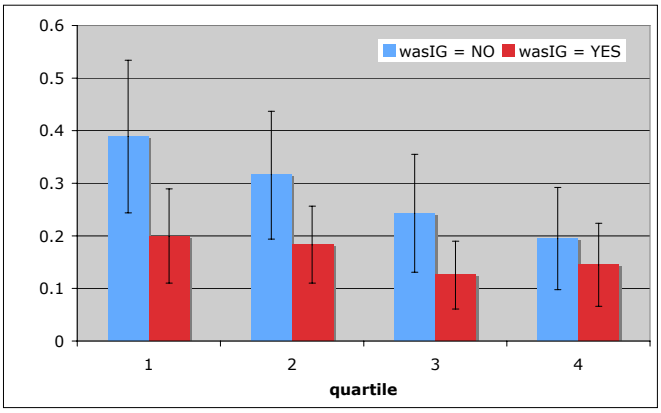

Figure 6: Interaction between experience and quartile

Useful features The rate of the useful features significantly increased within dialogues ($F_1(3, 93) = 5.71$, $p = 0.001$). The effect is mainly due to the increase of use of feature terms in the last quartile.

In 82.5% of introductory mentions, the distinguishing feature is mentioned, cf. Figure 7 and Table 1. The pattern maps are a special case, because the maps were unintentionally designed in a way that number is a reliable predictor, simply because the landmarks were visually grouped to a much higher degree than in the other maps. Pattern and number terms taken together, however, are used in 83.7% of introductory mentions.

Although there is no obvious reason why useful features should increase in the last quartile in particular, the trend shows that the contrasting fall in the use of color terms cannot be due to a general decrease in the feature mentions over the course of a dialogue.

Table 1: Means of feature terms used in quartiles (values for pattern maps are the sum of pattern and number terms; cf. text)

	Q1	Q2	Q3	Q4
Color	0.337	0.293	0.213	0.216
Number	0.871	0.833	0.845	0.939
Pattern	0.746	0.870	0.824	0.909
Kind	0.809	0.777	0.671	0.888
Shape	0.807	0.819	0.819	0.824

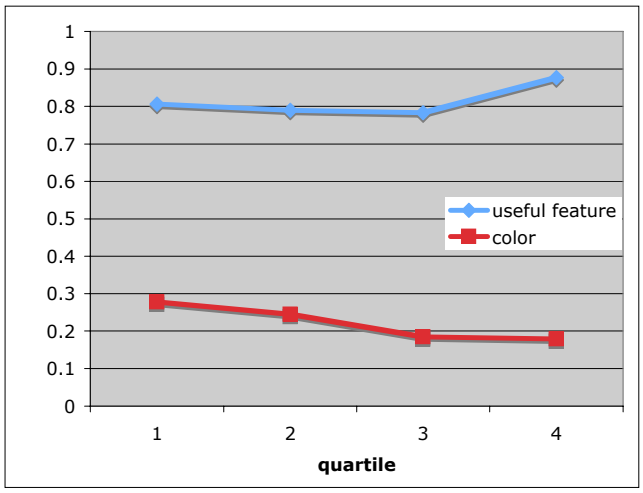

Figure 7: Change in the use of useful features and color within dialogues.

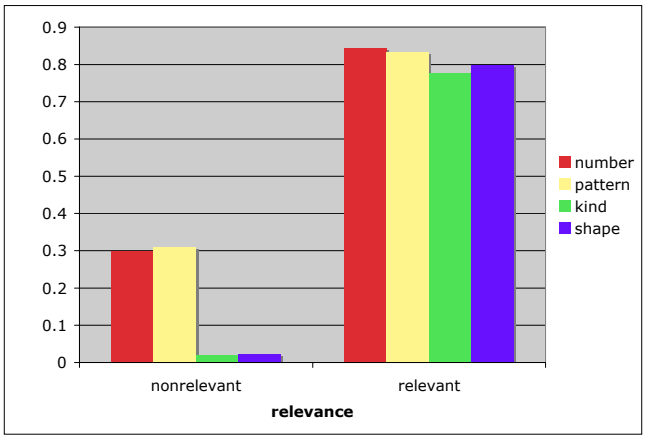

Figure 8: Relevance of feature terms in relevant and non-relevant maps

Relevance of feature terms

To establish whether feature terms are used more for the maps for which they are the distinguishing feature, we compared these two dialogues to one randomly selected pair of dialogues in which this feature was not distinguishing (see Figure 8). All four features (number, pattern, shape, kind) occur at a significantly higher rate in the maps for which they are distinguishing than in the control maps (number $F_1(1, 31) = 224.97$, $p < 0.001$; pattern $F_1(1,30) = 57.41$, $p < 0.001$; kind $F_1(1, 29) = 304.09$, $p < 0.001$; shape $F_1(1, 29) = 665.89$, $p < 0.001$).

Discussion

In our data, we find that the participants adapt to the properties of the stimulus maps they are presented with. The strongest effect is for the relevance of the feature terms. That is, the participants correctly maintain the right distinguishing feature for each map, without maintaining features that are possible but not critical. This is perhaps not surprising and should be covered by all existing algorithms for the generation of referring expressions.

In addition to this, however, we also find that the participants adapt to the utility of the features. The first adaptation is a global one to the fact that color is an unreliable distinguisher, because it cannot only mismatch between the maps for individual landmarks but is actually obscured for large portions of the Instruction Follower's map. Consequently, the use of color terms decreases. However, it is not a steady decline but a sharp drop when the participants swap the roles of Instruction Giver and Instruction Follower. This supports the view that speakers take a speaker-oriented approach: Once a player has been Instruction Follower and found that color is unreliable, he/she uses this feature significantly less often than the original Instruction Giver. Apparently, the initial salience of color is outweighed by its lack of utility here.

The second adaptation is a local one that affects the use of features that are tuned to the needs of the individual maps. We find a significant increase of the use of useful features (those with high utility) and a decrease in the feature with low utility – color.

Our utility-based explanation is corroborated by a simple computational cognitive model within the ACT-R theory that accounts for the changes in the use of color terms. We present this model in Guhe and Bard (2008).

Future research

We will continue analyzing this data on the level of individual landmark introductions to find out whether the adaptation to the relevant feature terms is decided locally for each single referring expression or whether it is global for an entire map. We will also analyze the use of spatial relations in landmark introductions, e.g. *The owl next to the purple penguin*.

Acknowledgements

This research was supported by grant NSF-IIS-0416128 to Max Louwerse, Art Graesser, Mark Steedman, and Ellen Gurman Bard. Thanks to Antje van Oosten and Jonathan Kilgour for help with the coding and data extraction and the Memphis team for providing transcriptions and pictures: Max Louwerse, Gwyneth Lewis and Megan Zirnstein.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, 34(4), 351–366.

Arnold, J. E. and Griffin, Z. M. (2007). The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, 56(4), 521–536.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M. P., Doherty-Sneddon, G., and Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, 42(1), 1–22.

Clark, H. H. (1996). *Using Language*. Cambridge University Press, Cambridge, MA.

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

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

Guhe, M. and Bard, E. G. (2007). Adaptation of the use of colour terms in referring expressions. In Proceedings of Decalog: The 11th Workshop on the Semantics and Pragmatics of Dialogue, University of Trento, May 30 – June 1, 2007, pages 167–168.

Guhe, M. and Bard, E. G. (2008) Adapting the use of attributes to the task environment in joint action: Results and a model. In: *Proceedings of Londial – The 11th Workshop on the Semantics and Pragmatics of Dialogue*.

Haywood, S. L., Pickering, M. J., and Branigan, H. P. (2005). Do speakers avoid ambiguities during dialogue? *Psychological Science*, 16(5), 362–366.

Jaeger, T. F. (2007) *Better categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models*. Manuscript under revision.

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

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

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics* 27, 98–110.

van der Sluis, I. (2005). *Multimodal Reference: Studies in Automatic Generation of Multimodal Referring Expressions*. PhD thesis, Tilburg University, The Netherlands.