# Referential choice in identification and route directions

**Adriana Baltaretu (a.a.baltaretu@tilburguniversity.edu)**

**Emiel J. Krahmer (e.j.krahmer@tilburguniveristy.edu)**

**Alfons A. Maes (maes@tilburguniversity.edu)**
Tilburg center for Cognition and Communication, PO Box 90153, 5000 LE, Tilburg, The Netherlands

**Abstract**

Though communicative goals are an important element in language production, few studies investigate the extent to which these goals might affect the form and content of referring expressions. In this study, we directly contrast two tasks with different goals: identification and instruction giving. Speakers had to refer to a target building nearby or further away, so that their addressee would distinguish it between other buildings (identification) or give route directions and use the same building as a landmark (instructions). Our results showed that irrespective of goals, the referring expressions consisted of the same types of attributes, yet the attribute frequency and formulation differed. In the identification task, references were longer, contained more locative and more post-nominal modifiers. In addition, referential choices were influenced by the visual distance between the speaker and the target: when the speaker observed the target from far, references were longer and contained more often locative modifiers.

**Keywords:** communicative goals; identification task; route directions; referring expressions.

Suppose you want to point out a building to a tourist, either because that is the hotel he is looking for or because it is part of the route direction you are asked to give. In both cases, you would have to describe the building (the target) in such a way that your addressee can distinguish it from the other buildings (the distractors). Most probably you will have to choose between different attributes (modifiers) that single out the building (e.g., color, location, size, architectural style). Though you need to refer to the same building, the two situations are rather different. In the first case, the goal is to help your addressee distinguish the target from similar objects (e.g., look at X). This is similar to an identification (or discrimination) task where the speaker has to utter a distinguishing description in order to identify a target (Van Deemter, Gatt, van Gompel, & Krahmer, 2012). Compared to an identification task, focused on describing the target, when giving instructions, the distinguishing description is part of an action oriented speech act (e.g., go to X and turn left). The main goal of the latter is helping your addressee to turn on the correct street. The question that arises is to what extent humans tune their referential choices when having different communicative goals (e.g., object identification, route directions)?

This paper focuses on referring expression production in naturalistic scenes, e.g., how a speaker chooses among different attributes and how his choice is influenced by communicative goals. We define route directions as an action oriented discourse composed of descriptive (references to objects) and procedural information (references to actions), aimed at helping a person navigate in an unknown environment (e.g., Allen, 1997; Michon & Denis, 2001). Procedural information refers to a series of actions that one needs to take, while descriptive information typically refers to environmental features (landmarks), such as buildings and their attributes (Allen, 1997). Paired with actions (e.g., turn at X), landmarks ground the direction change that has to be performed in the intersection (Michon & Denis, 2001) and positively affect the quality of the instructions and navigation performance (Tom & Tversky, 2012). By an identification task we refer to a situation in which the speaker has to produce a unique referring expression. The referring expression is required to be a description consisting of a set of properties that singles out a target from similar objects (Van Deemter et al., 2012). The focus of this study is on initial (definite) references whose content is shaped by information available in the visual physical context. In general, these target descriptions consist of a definite article, a head noun, and nominal modifiers.

## Communicative goals and referring expressions

The speaker's communicative goals are an important element of language production. This observation has been often acknowledged by qualitative work on dialogue development and some experimental work on reference and spatial language (Di Eugenio, Jordan, Thomason, & Moore, 2000; Daniel & Denis, 2004; Vorwerg & Tenbrink, 2006). Yet, as far as we know, there are no systematic comparisons assessing how much different referring expressions could be. We propose to investigate the extent to which communicative goals affect reference production, in a study that uses naturalistic scenes (depicting buildings in intersections), while taking into account a perceptual factor present in natural settings (the distance from which the target object is being observed). We compare the referring expressions produced when 'identifying' buildings, with the ones produced for the same objects while giving 'route directions', and assess how these references are evaluated by addressees.

There are several studies on reference that emphasize identification as a goal in itself (for a review, see Krahmer & Van Deemter, 2012; Van Deemter et al., 2012). These 'identification studies' start from a single goal and manipulate the attributes needed to identify objects (e.g., the STARS corpus (Paraboni, Galindo, & Iacovelli, 2016); the TUNA corpus (van Deemter, van der Sluis, & Gatt, 2006); the GRE3D3 corpus (Viethen & Dale, 2008, 2011). These studies focus on the properties of the target in contrast to other objects presented in a simple (sometimes grid-like) visual context. The speaker's purpose is to get the addressee pick out the target

object with a particular description. In general, identification is assumed to be part of a larger cooperative interaction (e.g., Brennan & Clark, 1996). However, this interaction is rarely modelled, and this type of studies typically report on a narrow set of experimental tasks. This raises the question to what extent these results would generalize when speakers have different goals. Do speakers produce similar references when describing an object for their addressee and when giving route directions, or do they adapt their references to the communicative context? How would the perceptual complex nature of the scenes influence reference production?

There are reasons to believe that different goals might affect referential choices. For example, when the participants' goal was to negotiate, rather than just identify, the preference for different attributes changed (Di Eugenio et al., 2000). In the COCONUT corpus, players had to buy items on a fixed budget, and in order to carry out this task, they had to describe and negotiate the furniture items that they believed were relevant to the task. In doing so, the item's price became one of the most often used attribute. Moreover, different goals may result in references with different levels of specification (Yoon, Koh, & Brown-Schmidt, 2012). When speakers had to give instructions they avoided scalar adjectives that were unnecessary for identification, and when they had to describe events, they referred more frequently to the objects' size.

Not only communicative goals, but also the importance attached to them might influence the speaker's reference production. When speakers had to identify surgeon' 'buttons on a control panel, compared to click on 'elements for another participant, they were more likely to include detailed descriptions containing redundant attributes and location information (Arts, Maes, Noordman, & Jansen, 2011). These had ultimately helped the addressee fulfil his task faster. Similarly, when participants had the goal of instructing someone how to operate once an alarm clock, compared to teaching someone that needs remember how to do this every night, their referring expressions contained more detailed information (Maes, Arts, & Noordman, 2004).

In everyday language, ones' communicative goals are realized through choices among the expressive means (lexical items, omissions/deletions, syntactic structure, word order, phrases and prepositions, among others; Talmy (2000, p. 346). In this study, we directly compare two tasks with different communicative goals using the same set of visual scenes. It is conceivable that references might have a different level of specification and the types of attributes and their distribution might be different across the two tasks. For example, as speakers focus more on the target in the identification task, they might mention more details about the buildings (such as the number of windows on a façade). On the other hand, route directions have a strong instructional focus and procedural information is arguably more important for the success of the task. Though, referring to a landmark also requires an 'identification' step, it might be more important what the speaker is trying to get the addressee to do with the utterance. As speakers need to convey two types of information, we might expect shorter and more focused references in the route directions than in the identification task.

Apart from communicative goals, a second factor that we explore in relation to reference production and comprehension is related to the distance from which the participants perceive the target. A difference of distance would affect the size of the target, the amount of visual details, and the number of objects in the visual field. In such situations, perception and recognition of object properties might be harder to assess for both speaker and listener; for example, the visibility of some object (or of one of its parts) may not be inferred with complete certainty. Distance differences seem to trigger various strategies when producing references. For example, participants tend to point more often when close to the object and refer less to the target's location, but use more locative phrases when pointing becomes ambiguous (Van Der Sluis & Krahmer, 2004; Bangerter, 2004). We question when pointing is not an available option, how would references change. When the intersection is far, the references might be longer, due to high ambiguity levels, however the opposite might be the case as from close by one has access to more visual details.

# Methods

## Participants

Eighty native Dutch-speaking students (40 dyads) of Tilburg University (62 women, mean age 21 years) participated in exchange for partial course credits. Participants were randomly assigned to speaker roles (28 women). The study was carried out in accordance with the recommendations of APA guidelines for conducting experiments, and all participants gave written consent for use of their data.

## Materials

Experimental materials consisted of 36 target objects (buildings depicted in Google Street View snapshots of intersections). These buildings were pictured from two distance points (36 scenes taken with a camera positioned at 40m away from the target and 36 scenes taken with a camera positioned at 20m away from the target). This resulted in 72 experimental scenes. The target objects were buildings marked with red squares, and placed in the corners of 4-way intersections (see Figure 1). Targets were always placed in the corners with highest visibility, namely on the other side the intersection. The targets' position was counterbalanced, so that in half of the scenes they were placed on the left side of the intersection. In both tasks we used the same scenes. For the route directions task, we added arrows indicating the turning direction (see Figure 2).

In addition, 36 filler scenes were introduced in order to present participants with a range of different navigation scenarios, and avoid participants from relying on fixed responses. The filler scenes depicted buildings in intersections with complex geometric structures.

## Procedure

Participants worked in pairs, and completed their task on separate computers. Pointing was discouraged by placing the computer screens in between the participants. The speaker

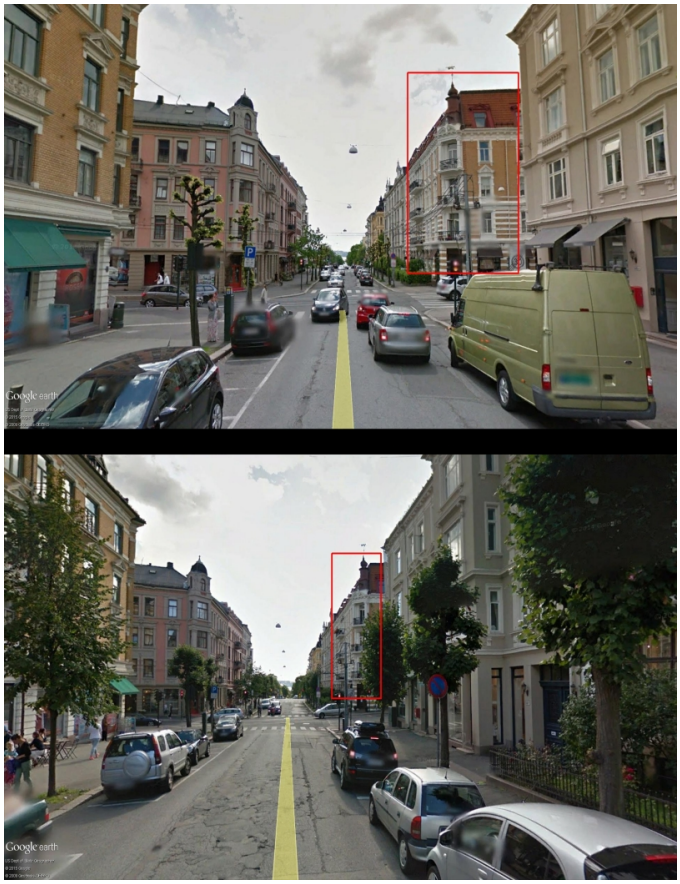

Figure 1: Experimental scenes from the description task depicting a target building near (above) and far (below)

Figure 2: Experimental scenes from the route direction task depicting a target building near (above) and far (below)

received a scene with a marked building (and an arrow indicating the route), while the listener received the same image without any markings.

Participants were randomly assigned to one of the two tasks. In the descriptions task (D-task), speakers were asked to refer to the marked building in such a way that the listener could uniquely identify it. In the route directions task (RD-task), speakers were asked to give route directions, and in doing so make use of the marked building as landmark. In order to elicit uniform responses, speakers were asked to verbally fill in templates (a typical procedure in identification studies, e.g., Dale & Viethen, 2009). They had to fill in the following templates: "click on ... " (descriptions) and "go to ... and turn left / right" (route directions). In both situations, the listeners had to click on the correct building. Listeners were allowed to ask questions if the speakers' instructions were unclear. Experimental scenes were divided in presentation lists, so that each participant would see each target object only once. Participants were randomly assigned to one of the presentation lists. The experiment started with three practice trials, next 72 trials (36 experimental trials and 36 filler trials) were presented in different random orders.

## Design

This study had a 2 x 2 design with Task (levels: D-task, RD-task) as between participants factor and Distance (levels: far, near) as within participants factor. For the first analysis, we looked at the length of the references (number of words). Next, we assessed the type of attributes mentioned in relation with the target (e.g., color, location, etc), their frequency and distribution (pre / post nominal modifiers), as well as the length of the remaining phrase after the speaker has mentioned the noun denoting the target building (number of words). The presence / absence of attributes coupled with the target noun was binary coded. A phrase like *the large building on the left, next to the white tower* consists of a target noun (*building*) and the following attributes: size (*large*) and location (*on the left, next to the white building*). In order to analyze the differences regarding the length of the references, data transformations (log data) were applied due to a skewed distribution. For ease of understanding, means and standard deviations reported here represent untransformed data.

In order to test the observed differences, we conducted separate statistical analyses using logit mixed model analysis (Jaeger, 2008), following the recommendations of Barr, Levy, Scheepers, and Tily (2013). We used the mixed logit model

analysis as it can correctly account for random subject and item effects in a one-step analysis. The models were fitted using the LMER function from the LanguageR Package in R (version 2.15.2; CRAN project; The R Foundation for Statistical Computing, 2012). To determine whether the two conditions significantly differed from each other, we started by constructing a maximal model with a full random effect structure. This had Task and Distance as fixed factors; Speakers and Scenes as random factors; intercepts and random slopes for Speakers and Scenes to account for between-subject and between-item variation. In case the dependent variable was binary coded, the factors were centered to avoid collinearity. In case the model did not converge, we only excluded random slopes with the lowest variance until convergence was reached. The results from the first converging model, as well as the structure of the model are reported. The $p$ - values were estimated via parametric bootstrapping over 100 iterations.

## Results

In total 1440 references were produced (40 speakers * 36 scenes). Data from two dyads (one from each condition) were discarded on the basis of task misunderstanding. The referring expressions consisted of a noun denoting the target object and all the phrases attached to it. In practice, in the D-task, the references consist of all the phrases after "click on", and in the RD-task, they consist of all the phrases between "Go to" and "and turn". In some cases, participants omitted some part of the template, but kept the overall structure (e.g., *The red building, and turn ...*), and these cases were included in the analysis. In fact, 91 percent of the references had a similar structure (noun and modifiers). The remaining, nine percent of the cases resulted in utterances with different structure (e.g., *The building is ...*), evenly distributed across the two tasks (5% cases in the D-task and 4% in the RD-task). As this could potentially bias aspects such as word counts, we decided to exclude these cases from the analysis.

### Length of referring expressions

The first converging model had random intercepts for Speakers and Scenes, $R^2$ marginal = 0.10, $R^2$ conditional = 0.56. The length of the references was significantly influenced by the Task ($\beta = -0.369$; $SE = 0.12$; $p < .01$). Referring expressions were longer in the D-task ($M = 16.3$, $SD = 1.3$), than in the RD-task ($M = 11$, $SD = 1.3$). There was a main effect of Distance ($\beta = 0.076$; $SE = 0.03$; $p < .01$). When close to the target, speakers produced slightly shorter referring expressions ($M = 13$, $SD = .88$), than when further away ($M = 14$, $SD = 1.01$). There was no interaction between Distance and Task ($p > .05$).

### Type of attributes

Speakers described targets by referring to ten types of attributes (see Table 1). The same types were produced in both tasks. Top three most frequent attributes in both tasks are location, followed by colour and references to structural parts of the target (such as chimney, stairs, doors).

In the D-task, speakers mentioned more often the location and color of the object than in the RD-task. We analyzed statistically these differences. With respect to location, the first converging model had random intercepts for Speakers and Scenes. There was a significant difference between the tasks regarding locative information ($\beta = -2.573$; $SE = 0.81$; $p < .01$). There was a main effect of Distance ($\beta = 0.535$; $SE = 0.22$; $p < 0.01$). When close to the target, speakers referred less often to the position of the object ($M = .77$, $SD = 0.41$), than when further away ($M = .82$, $SD = 0.32$). There was no significant interaction between Task and Distance ($p > .05$).

As for the difference regarding references to the target color, the first converging model had random intercepts for Speakers and Scenes, and there were no main effects ($p > .05$) or interactions ($p > .05$).

Table 1: Type of attributes, examples and attribute frequency split by task

| Type | Examples | Frequency | |
|---|---|---|---|
| | | Route Directions | Descriptions |
| location | The building across the intersection, on the left | 68% | 92% |
| color | The white building | 42% | 47% |
| part of building | The building with (five) balconies / with (red) roof / with (two) windows | 35% | 36% |
| decoration | The building with stripes / with flowers pots / with hanging things / with flags | 11% | 12% |
| size | The smallest building | 8% | 7% |
| shape | The long building | 0.6% | 0.3% |
| age | The modern building | 1% | 0.5% |
| architectural style | The Italian building | 2% | 0.3% |
| materials | The brick building | 1% | 1% |
| evaluative | The ugly building | 1% | 2% |

### Distribution of modifiers

First there was significant difference regarding the number of words produced after the target noun and the first converging model had random intercepts for Speakers and Scenes ($\beta = -0.508$, $SE = 0.16$, $p < .01$). The number of words produced after the target noun was longer in the D-task ($M = 13.62$, $SD = 1.25$), than in the RD-task ($M = 7.81$, $SD = 1.25$). There was no main effect of Distance ($p > .05$) and no interaction between the factors ($p > .05$).

In general, there were more post-nominal modifiers in the D-task ($N = 1002$) compared to the RD-task ($N = 683$). The reversed pattern was observed for pre-nominal modifiers, which were more frequent in the RD-task ($N = 306$) than in the D-task ($N = 212$). In both tasks, pre-nominal modifiers consisted mostly of location (e.g., *the left building*; *the second building*), color (e.g., *the white building*) and size references (e.g., *the large building*) (see Table 2). Post-nominal modifiers consisted also of location, color and size, but mostly included references to structural parts of the building that syntactically can not be framed otherwise (e.g., *the building with two balconies*).

Table 2: Distribution (number of cases) of pre- and post-nominal modifiers split by task.

| | Pre-nominal | | Post-nominal | |
|---|---|---|---|---|
| | Route Directions | Descriptions | Route directions | Descriptions |
| location | 151 | 116 | 270 | 434 |
| colour | 128 | 70 | 136 | 221 |
| size | 21 | 17 | 27 | 28 |
| other | 6 | 9 | 250 | 319 |

### Lexical Fillers

We define lexical fillers (hedges) as words or phrases that are conventionally used for signalling hesitation, marking the reference as provisional (Brennan & Clark, 1996). The references produced in the two tasks included a different amount of lexical fillers (see Figure 3). D-task triggered more lexical fillers, compared to the RD-task.

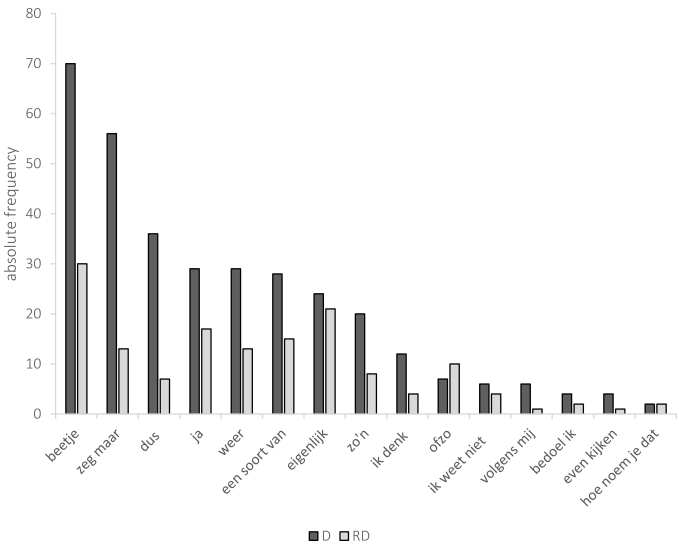


Figure 3: Distribution (number of cases) of lexical fillers split by task

### Error rates

There was a small number of cases in which the addressee clicked on wrong buildings (3 cases in RD-task and 6 in the D-task) and relatively few clarification questions (13 questions in the D-task and 26 questions in the RD-task). These questions mostly asked for simple clarifications (e.g., *a big building?*) to which speakers mostly uttered a short confirmation (e.g., *yes*).

## Conclusions and Discussion

In this study, we questioned to what extent different communicative goals (object identification, giving route directions) influence reference production, while using complex naturalistic scenes and taking into account a perceptual factor, the visual distance from which a target object is observed. The referring expressions were elicited using real-world scenes, and the design allowed a direct comparison between the references elicited in two communicative settings. Preliminary results showed that identification as opposed to referring to objects in route directions triggered a number of differences regarding reference formulation, and almost no semantic differences.

First, there were no semantic differences, in both tasks speakers used the same types of attributes to refer to buildings. This pattern of results suggests that studies, where identification is the main purpose of the interaction, could generalize to a large extent to other settings. Second, the syntactic formulation of the references was different. Descriptions were longer and contained more post-nominal information and more words following the target noun than references in route directions, which suggests that speakers described the target in more detail. Contrastively, in route directions, referring expressions were shorter and tended to contain more pre-nominal modifiers. Previous research suggests that attributes in pre-nominal position are more efficient for identification (Rubio-Fernndez, 2016), which might suggest that in a larger communicative act speakers 'optimize' their references, in such a way that the addressee can find the target faster and more easily (see also, Clarke, Elsner, & Rohde, 2015). We plan to test how differences in information structure affect addressee performance, when one needs to find his way in an unknown environment. Moreover, in the description task, speakers tended to pay more attention to the location of the object. Locating an entity has been suggested to be a robust and successful strategy in object identification (Paraboni et al., 2016) though it could have contributed to the increased length of the descriptions (see also Vorwerg & Tenbrink, 2006). In addition, and unexpectedly, on the pragmatic level, we found different levels in the use of markers of nuance and hesitation. We question whether the difference in absolute length between tasks was caused by the post-nominal structures, the pragmatic hesitations, or more words to express some attributes as with location? We assume it is a combination of these three.

These differences could be caused by the type of discourse elicited in the tasks. On the one hand, action related information might be more important for route directions. The difference in length might be explained by the fact that speakers want to make sure the addressee finds the correct street, and less so the correct building. Moreover, landmarks are meant to improve route directions, and long complex references would doubtfully be an efficient contribution. On the other hand, descriptions might be a more difficult and less planned task, as suggested by a higher level of post-nominal information and lexical fillers.

In addition, we explored if the distance between the speaker and the target might influence referring expressions. This factor relates to several aspects that increase the uncertainty with respect to what building is referred to. For example when the speaker is further from the target, there are more similar buildings in the scene and the target could be partially occluded. References were longer and contained more location information when the speaker was far from the target.

However the differences between the two conditions, as well as the effect size, were rather small.

In sum, when comparing the effects of different communicative goals on reference production an interesting pattern of results emerges. In line with studies focused exclusively on identification, we noticed a stable preference for the same type of attributes. Yet, different communicative goals seem to shape the information structure of the references and influence the frequency of some of the attributes.

## References


Allen, G. L. (1997). From knowledge to words to wayfinding: Issues in the production and comprehension of route directions. In S. C. Hirtle & A. U. Frank (Eds.), *Spatial Information Theory A Theoretical Basis for GIS* (pp. 363–372). Berlin: Springer.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*(1), 361–374.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, *15*(6), 415–419.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482-1493.

Clarke, A. D. F., Elsner, M., & Rohde, H. (2015). Giving good directions: order of mention reflects visual salience. *Frontiers in Psychology*, *6*(1793). doi: 10.3389/fpsyg.2015.01793

Dale, R., & Viethen, J. (2009). Referring expression generation through attribute-based heuristics. In E. Krahmer & M. Theune (Eds.), *Proceedings of the 12th european workshop on natural language generation* (pp. 58–65). Stroudsburg, PA, USA: Association for Computational Linguistics.

Daniel, M.-P., & Denis, M. (2004). The production of route directions: Investigating conditions that favour conciseness in spatial discourse. *Applied Cognitive Psychology*, *18*(1), 57–75.

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

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

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

Maes, A., Arts, A., & Noordman, L. (2004). Reference management in instructive discourse. *Discourse Processes*, *37*(2), 117–144.

Michon, P.-E., & Denis, M. (2001). When and why are visual landmarks used in giving directions? In D. R. Montello (Ed.), *Spatial information theory* (pp. 292–305). Berlin: Springer.

Paraboni, I., Galindo, M. R., & Iacovelli, D. (2016). Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*, 1–24.

Rubio-Fernndez, P. (2016). How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in Psychology*, *7*(153). doi: 10.3389/fpsyg.2016.00153

Talmy, L. (2000). *Toward a cognitive semantics* (Vol. 1). MIT press.

Tom, A. C., & Tversky, B. (2012). Remembering routes: streets and landmarks. *Applied Cognitive Psychology*, *26*(2), 182–193.

Van Deemter, K., Gatt, A., van Gompel, R. P., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, *4*(2), 166–183.

van Deemter, K., van der Sluis, I., & Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the fourth international natural language generation conference* (pp. 130–132). Stroudsburg, PA, USA: Association for Computational Linguistics.

Van Der Sluis, I., & Krahmer, E. (2004). The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In S. Kim & D. Youn (Eds.), *Proceedings of international conference on spoken language processing* (p. 1005-1008). Korea: ISCA.

Viethen, J., & Dale, R. (2008). The use of spatial relations in referring expression generation. In W. Michael, N. Crystal, & M. David (Eds.), *Proceedings of the fifth international natural language generation conference* (pp. 59–67). Stroudsburg, PA, USA: Association for Computational Linguistics.

Viethen, J., & Dale, R. (2011). Gre3d7: A corpus of distinguishing descriptions for objects in visual scenes. In B. Anja, E. Roger, G. Albert, & S. Kristina (Eds.), *Proceedings of the ucnlg and evaluation: Language generation and evaluation workshop* (pp. 12–22). Stroudsburg, PA, USA: Association for Computational Linguistics.

Vorwerg, C., & Tenbrink, T. (2006). Discourse factors influencing spatial descriptions in english and german. In T. Barkowsky, M. Knauff, G. Ligozat, & D. Montello (Eds.), *Spatial Cognition V Reasoning, Action, Interaction* (pp. 470–488). Berlin Heidelberg: Springer-Verlag.

Yoon, S. O., Koh, S., & Brown-Schmidt, S. (2012). Influence of perspective and goals on reference production in conversation. *Psychonomic Bulletin & Review*, *19*(4), 699–707.