# Anchor-Progression in Spatially Situated Discourse: a Production Experiment

**Hendrik Zender** and **Christopher Koppermann** and **Fai Greeve** and **Geert-Jan M. Kruijff**
Language Technology Lab
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
zender@dfki.de

## Abstract

The paper presents two models for producing and understanding situationally appropriate referring expressions (REs) during a discourse about large-scale space. The models are evaluated against an empirical production experiment.

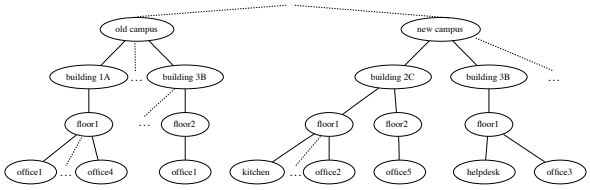
## 1 Introduction and Background

For situated interaction, an intelligent system needs methods for relating entities in the world, its representation of the world, and the natural language references exchanged with its user. Human natural language processing and algorithmic approaches alike have been extensively studied for application domains restricted to small visual scenes and other small-scale surroundings. Still, rather little research has addressed the specific issues involved in establishing reference to entities outside the currently visible scene. The challenge that we address here is how the focus of attention can shift over the course of a discourse if the domain is larger than the currently visible scene.

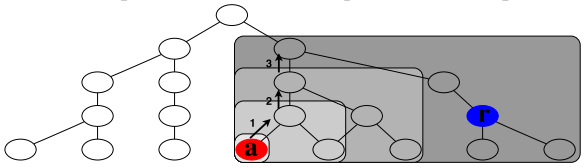
The generation of referring expressions (GRE) has been viewed as an isolated problem, focussing on efficient algorithms for determining which information from the domain must be incorporated in a noun phrase (NP) such that this NP allows the hearer to optimally understand which referent is meant. The domains of such approaches usually consist of small, static domains or simple visual scenes. In their seminal work Dale and Reiter (1995) present the Incremental Algorithm (IA) for GRE. Recent extensions address some of its shortcomings, such as negated and disjoined properties (van Deemter, 2002) and an account of salience for generating contextually appropriate shorter REs (Krahmer and Theune, 2002). Other, alternative GRE algorithms exist (Horacek, 1997; Bateman, 1999; Krahmer et al., 2003). However, all these algorithms rely on a given *domain of discourse* constituting the current *context* (or *focus of attention*). The task of the GRE algorithm is then to single out the intended referent against the other members of the context, which act as *potential distractors*. As long as the domains are such closed-context scenarios, the intended referent is always in the current focus. We address the challenge of producing and understanding of references to entities that are outside the current focus of attention, because they have not been mentioned yet and are beyond the currently observable scene.

Our approach relies on the dichotomy between *small-scale space* and *large-scale space* for human spatial cognition. Large-scale space is "a space which cannot be perceived at once; its global structure must be derived from local observations over time" (Kuipers, 1977). In everyday situations, an office environment, one's house, or a university campus are large-scale spaces. A table-top or a part of an office are examples of small-scale space. Despite large-scale space being not fully observable, people can nevertheless have a reasonably complete mental representation of, e.g., their domestic or work environments in their *cognitive maps*. Details might be missing, and people might be uncertain about particular things and states of affairs that are known to change frequently. Still, people regularly engage in a conversation about such an environment, making successful references to spatially located entities.

It is generally assumed that humans adopt a *partially hierarchical* representation of spatial organization (Stevens and Coupe, 1978; McNamara, 1986). The basic units of such a representation are *topological* regions (i.e., more or less clearly bounded spatial areas) (Hirtle and Jonides, 1985). Paraboni et al. (2007) are among the few to address the issue of generating references to entities outside the immediate environment, and present an algorithm for *context determination* in hierar-




(a) Example for a hierarchical representation of space.




(b) Illustration of the TA principle: starting from the attentional anchor ($a$), the smallest sub-hierarchy containing both $a$ and the intended referent ($r$) is formed incrementally.

Figure 1: TA in a spatial hierarchy.

chically ordered domains. However, since it is mainly targeted at producing textual references to entities in written documents (e.g., figures and tables in book chapters), they do not address the challenges of physical and perceptual situatedness. Large-scale space can be viewed as a hierarchically ordered domain. To keep track of the referential context in such a domain, in our previous work we propose the principle of *topological abstraction* (TA, summarized in Fig. 1) for context extension (Zender et al., 2009a), similar to Ancestral Search (Paraboni et al., 2007). In (Zender et al., 2009b), we describe the integration of the approach in an NLP system for situated human-robot dialogues and present two algorithms instantiating the TA principle for GRE and resolving referring expressions (RRE), respectively. It relies on two parameters: the location of the *intended referent* $r$, and the *attentional anchor* $a$. As discussed in our previous works, for single utterances the anchor is the physical position where it is made (i.e., the *utterance situation* (Devlin, 2006)). Below, we propose models for attentional anchor-progression for longer discourses about large-scale space, and evaluate them against real-world data.

## 2 The Models

In order to account for the determination of the attentional anchor $a$, we propose a model called *anchor-progression* $A$. The model assumes that each *exophoric* reference[1] serves as *attentional anchor* for the subsequent reference. It is based on observations on "principles for anchoring resource situations" by Poesio (1993), where the expression of movement in the domain determines the updated current mutual focus of attention. $a$ and $r$ are then passed to the TA algorithm. Taking into account the verbal behavior observed in our experiment, we also propose a refined model of *anchor-resetting* $R$, where for each new turn (e.g., a new instruction), the anchor is re-set to the *utterance situation*. $R$ leads to the inclusion of navigational information for each first RE in a turn, thus reassuring the hearer of the focus of attention.

[1]This excludes pronouns as well as other descriptions that pick up an existing referent from the linguistic context.

## 3 The Experiment

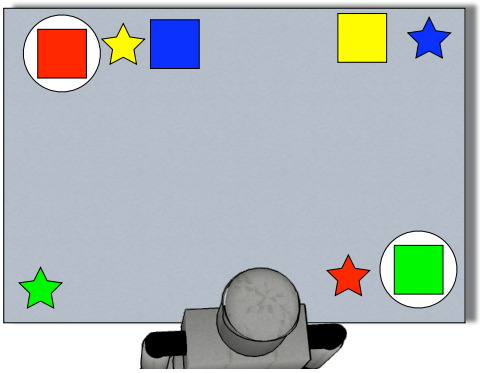
We are interested in the way the disambiguation strategies change when producing REs during a discourse about large-scale space versus discourse about small-scale space. In our experiment, we gathered a corpus of spoken instructions in two different situations: *small-scale space (SSS)* and *large-scale space (LSS)*. We use the data to evaluate the utility of the $A$ and $R$ models. We specifically evaluate them against the traditional (*global*) model $G$ in which the indented referent must be singled out from all entities in the domain.

The cover story for the experiment was to record spoken instructions to help improve a speech recognition system for robots. The participants were asked to imagine an intelligent service robot capable of understanding natural language and familiar with its environment. The task of the participants was to instruct the robot to clean up a working space, i.e., a table-top (SSS) and an indoor environment (LSS) by placing target objects (cookies or balls) in boxes of the same color. The use of color terms to identify objects was discouraged by telling the participants that the robot is unable to perceive color. The stimuli consisted of 8 corresponding scenes of the table-top and the domestic setting (cf. Fig. 2). In order to preclude the specific phenomena of collaborative, task-oriented dialogue (cf., e.g., (Garrod and Pickering, 2004)), the participants had to instruct an imaginary recipient of orders. The choice of a robot was made to rule out potential social implications when imagining, e.g., talking to a child, a butler, or a friend.

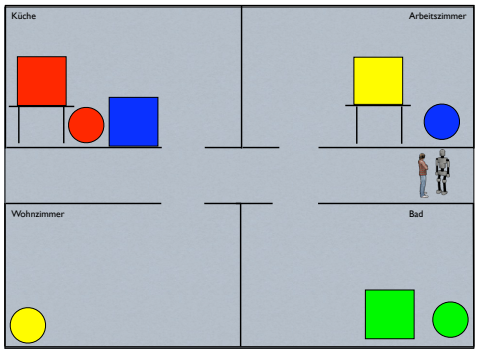
The SSS scenes show a bird's-eye view of the table including the robot's position (similar to (Funakoshi et al., 2004)). The way the objects are arranged allows to refer to their location with respect to the corners of the table, with plates as additional landmarks. The LSS scenes depict an indoor environment with a corridor and, parallel to SSS, four rooms with tables as landmarks. The scenes show

Table 1: Example from the small-scale (1–2) and large-scale space (3–4) scenes in Fig. 2.

1. *nimm [das plätzchen unten links]*$_{m_{G,A}}$ *, leg es [in die schachtel unten rechts auf dem teller]*$_{o_{G,A}}$
   'take the cookie on the bottom left, put it into the bottom right box on the plate'
2. *nimm [das plätzchen unten rechts]*$_{m_G,o_A}$ *, leg es [in die schachtel oben links auf dem teller]*$_{m_{G,A}}$
   'take the cookie on the bottom right, put it into the top left box on the plate'
3. *geh [ins wohnzimmer]*$_{m_{G,A,R}}$ *und nimm [den ball]*$_{u_G,m_{A,R}}$ *und bring ihn [ins arbeitszimmer]*$_{m_{G,A,R}}$ *, leg ihn [in die kiste auf dem tisch]*$_{u_G,o_{A,R}}$
   'go to the living room and take the ball and bring it to the study; put it into the box on the table'
4. *und nimm [den ball]*$_{u_{G,R},m_A}$ *und bring ihn [in die küche]*$_{m_{G,A,R}}$ *und leg ihn [in die kiste auf dem boden]*$_{u_G,m_{A,R}}$
   'and take the ball and bring it to the kitchen and put it into the box on the floor'



(a) Small-scale space: squares represent small boxes, stars cookies, and white circles plates.




(b) Large-scale space: squares represent boxes placed on the floor or on a table, circles represent balls, rooms are labeled.

Figure 2: Two stimuli scenes from the experiment.

the robot and the participant in the corridor.

In order to gather more comparable data we opted for a *within-participants* approach. Each person participated in the *SSS treatment* and in the *LSS treatment*. To counterbalance potential carry-over effects, half of the participants were shown the treatments in inverse order, and the sequence of the 8 scenes in each treatment was varied in a principled way. In order to make the participants produce multi-utterance discourses, they were required to refer to all target object pairs. The exact wording of their instructions was up to them.

Participants were placed in front of a screen and a microphone into which they spoke their orders to the imaginary robot, followed by a self-paced keyword after which the experimenter showed the next scene. The experiment was conducted in German and consisted of a pilot study (10 participants) and the main part (19 female and 14 male students, aged 19–53, German native speakers). The data of three participants who did not behave according to the instructions was discarded. The individual sessions took 20–35 min., and the participants were paid for their efforts.

Using the UAM CorpusTool software, transcriptions of the recorded spoken instructions were annotated for occurrences of the linguistic phenomenon we are interested in, i.e., REs. Samples were cross-checked by a second annotator. REs were marked as shallow 'refex' segments, i.e., complex NPs were not decomposed into their constituents. Only definite NPs representing exophoric REs (cf. Sec. 2) qualify as 'refex' segments. If a turn contained an indefinite NP, the whole turn was discarded. The 'refex' segments were coded according to the amount of information they contain, and under which disambiguation model $M \in \{G, A, R\}$ ($R$ only for LSS) they succeed in singling out the described referent. Following Engelhardt et al. (2006), we distinguish three types of semantic specificity. A RE is an *over-description* with respect to $M$ ($over_M$) if it contains redundant information, and it is an *under-description* ($under_M$) if it is ambiguous according to $M$. *Minimal descriptions* ($min_M$) contain just enough information to uniquely identify the referent. Table 1 shows annotated examples.

## 4 Results

The collected corpus consists of 30 annotated sessions with 2 treatments comprising 8 scenes with 4 turns. In total, it contains 4,589 annotated REs, out of which only 83 are errors. Except for the error rate calculation, we only consider non-error 'refex' segments as the universe. The SSS treat-

Table 2: Mean frequencies (with standard deviation in italics) of minimal ($min$), over-descriptions ($over$), and under-descriptions ($under$) with respect to the models ($A$, $R$, $G$) in both treatments.

| | $over_G$ | $over_A$ | $over_R$ | $min_G$ | $min_A$ | $min_R$ | $under_G$ | $under_A$ | $under_R$ |
|---|---|---|---|---|---|---|---|---|---|
| small-scale space | 13.94% *15.85%* | 34.45% *14.37%* | | 78.90% *17.66%* | 60.11% *13.13%* | | 7.16% *12.07%* | 5.43% *10.50%* | |
| large-scale space | 6.81% *7.53%* | 34.75% *12.13%* | 20.06 % *10.10%* | 68.04% *17.87%* | 64.55% *13.13%* | 76.73% *10.66%* | 25.16% *19.48%* | 0.69% *1.72%* | 3.21% *5.06%* |

ment contains 1,902 'refex', with a mean number of 63.4 and a std. dev. $\sigma$=1.98 per participant. This corresponds to the expected number of 64 REs to be uttered: 8 scenes $\times$ 4 target object pairs. The LSS treatment contains 2,604 'refex' with an average of 86.8 correct REs ($\sigma$=18.19) per participant. As can be seen in Table 1 (3–4), this difference is due to the participants' referring to intermediate waypoints in addition to the target objects. Table 2 summarizes the analysis of the annotated data.

Overall, the participants had no difficulties with the experiment. The mean error rates are low in both treatments: 1.78% ($\sigma$=3.36%) in SSS, and 1.80% ($\sigma$=2.98%) in LSS. A paired sample t-test of both scores for each participant shows that there is no significant difference between the error rates in the treatments ($p$=0.985), supporting the claim that both treatments were of equal difficulty. Moreover, a MANOVA shows no significant effect of treatment-order for the verbal behavior under study, ruling out potential carry-over effects.

Production experiments always exhibit a considerable variation between participants. When modeling natural language processing systems, one needs to take this into account. A GRE component should produce REs that are easy to understand, i.e., ambiguities should be avoided and over-descriptions should occur sparingly. A GRE algorithm will always try to produce minimal descriptions. The generation of an under-description means a failure to construct an identifying RE, while over-descriptions are usually the result of a globally 'bad' incremental construction of the generated REs (as is the case, e.g., in the IA). An RRE component, on the other hand, should be able to identify as many referents as possible by treating as few as possible REs as under-descriptions.

The analysis of the SSS data with respect to $G$ establishes the baseline for a comparison with other experiments and GRE approaches. 13.9% of the REs contain redundant information ($over_G$), compared to 21% in (Viethen and Dale, 2006). In contrast, however, our SSS scenes did not provide the possibility for producing more-than-minimal REs for every target object, which might account for the difference. $under_G$ REs occur with a frequency of 7.2% in the SSS data. Because under-descriptions result in the the hearer being unable to reliably resolve the reference, this means that the robot in our experiment cannot fulfill its task. This might explain the difference to the 16% observed in the task-independent study by Viethen and Dale (2006). The significantly ($p$<0.001) higher mean frequency of $min_G$ than $min_A$ underpins that $G$ is an accurate model for the verbal behavior in SSS. However, $G$ does not fit the LSS data well. An RRE algorithm with model $G$ would fail to resolve the intended referent in 1 out of 4 cases (cf. $under_G$ in LSS). With only 0.7% $under_A$ REs on average, $A$ models the LSS data significantly better ($p$<0.001). Still, there is is a high rate of $over_A$ REs. In comparison, $R$ yields a significantly ($p$<0.001) lower amount of $over_R$. The mean frequency of $under_R$ is significantly ($p$=0.010) higher than for $under_A$, but still below $under_G$ in the SSS data. With a mean frequency of 76.7% $min_R$, $R$ models the data better than both $G$ and $A$. For the REs in LSS $min_R$ is in the same range as $min_G$ for the REs in SSS.

## 5 Conclusions

Overall, the data exhibit a high mean frequency of over-descriptions. However, since this means that the human-produced REs contain more information than minimally necessary, this does not negatively affect the performance of an RRE algorithm. For a GRE algorithm, however, a more cautious approach might be desirable. In situated discourse about LSS, we thus suggest that $A$ is suitable for the RRE task because it yields the least amount of unresolvable under-descriptions. For the GRE task $R$ is more appropriate. It strikes a balance between producing short descriptions and supplementing navigational information.

## Acknowledgments

This work was supported by the EU Project CogX (FP7-ICT-215181). Thanks to Mick O'Donnell for his support with the UAM CorpusTool.

## References


John A. Bateman. 1999. Using aggregation for selecting content when generating referring expressions. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL'99)*, pages 127–134, Morristown, NJ, USA.

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

Keith Devlin. 2006. Situation theory and situation semantics. In Dov M. Gabbay and John Woods, editors, *Logic and the Modalities in the Twentieth Century*, volume 7 of *Handbook of the History of Logic*, pages 601–664. Elsevier.

Paul E. Engelhardt, Karl G.D. Bailey, and Fernanda Ferreira. 2006. Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54(4):554–573.

Kotaro Funakoshi, Satoru Watanabe, Naoko Kuriyama, and Takenobu Tokunaga. 2004. Generation of relative referring expressions based on perceptual grouping. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, Morristown, NJ, USA.

Simon Garrod and Martin J. Pickering. 2004. Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1):8–11, January.

Stephen C. Hirtle and John Jonides. 1985. Evidence for hierarchies in cognitive maps. *Memory and Cognition*, 13:208–217.

Helmut Horacek. 1997. An algorithm for generating referential descriptions with flexible interfaces. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL-97)*, pages 206–213, Morristown, NJ, USA.

Emiel Krahmer and Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and R. Kibble, editors, *Information Sharing: Givenness and Newness in Language Processing*, pages 223–264. CSLI Publications, Stanford, CA, USA.

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

Benjamin Kuipers. 1977. *Representing Knowledge of Large-scale Space*. PhD thesis, MIT-AI TR-418, Massachusetts Institute of Technology, Cambridge, MA, USA, May.

Timothy P. McNamara. 1986. Mental representations of spatial relations. *Cognitive Psychology*, 18:87–121.

Ivandré Paraboni, Kees van Deemter, and Judith Masthoff. 2007. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254, June.

Massimo Poesio. 1993. A situation-theoretic formalization of definite description interpretation in plan elaboration dialogues. In Peter Aczel, David Israel, Yasuhiro Katagiri, and Stanley Peters, editors, *Situation Theory and its Applications Volume 3*, CSLI Lecture Notes No. 37, pages 339–374. Center for the Study of Language and Information, Menlo Park, CA, USA.

Albert Stevens and Patty Coupe. 1978. Distortions in judged spatial relations. *Cognitive Psychology*, 10:422–437.

Kees van Deemter. 2002. Generating referring expressions: boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52.

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

Hendrik Zender, Geert-Jan M. Kruijff, and Ivana Kruijff-Korbayová. 2009a. A situated context model for resolution and generation of referring expressions. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 126–129, Athens, Greece, March.

Hendrik Zender, Geert-Jan M. Kruijff, and Ivana Kruijff-Korbayová. 2009b. Situated resolution and generation of spatial referring expressions for robotic assistants. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)*, pages 1604–1609, Pasadena, CA, USA, July.