Generating Underspecified Descriptions of Landmark Objects

Ivandré Paraboni, Alan K. Yamasaki, Adriano S. R. da Silva, and Caio V. M. Teixeira

School of Arts, Sciences and Humanities, University of São Paulo (EACH / USP)
Av. Arlindo Bettio, 1000 - São Paulo, Brazil
ivandre@usp.br

Abstract. We present an experiment to collect referring expressions produced by human speakers under conditions that favour landmark underspecification. The experiment shows that underspecified landmark descriptions are not only common but, under certain conditions, may be largely preferred over minimally and fully-specified descriptions alike.

Keywords: Natural Language Generation, Referring Expressions

1 Introduction

In Natural Language Generation (NLG), Referring Expression Generation (REG) [3,9] is the computational task of providing adequate referring expressions (e.g., pronouns, definite descriptions, proper names etc.) for discourse entities.

Let us consider the issue of selecting the semantic contents that a referring expression should convey.[1] For instance, suppose a domain containing two identical desks side-by-side, and a single man behind the right desk. The following are possible examples of uniquely identifying reference to the target man.

a The man
b The tall man
c The man behind the desk
d The tall man behind the desk
e The tall man behind the desk, on the right side

In this paper we will focus on the problem of generating non-ambiguous *relational* referring expression between a target r and a landmark object o, as in (c-e) above. More specifically, we will discuss the amount of information used by human speakers to produce the landmark (desk) description $L(o)$ as part of a larger reference to the main target (man).

We will say that the reference to a landmark object is *underspecified* when $L(o)$ does not fully distinguish o from all other objects in the same context, that is, when target and landmark descriptions are meant to mutually disambiguate each other as in (c) and (d). Conversely, we will say that $L(o)$ is *fully-specified* when $L(o)$ denotes o and no other object in the context as in (e).

[1] For surface realisation issues, see, e.g., [14,15,11].

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 76–83, 2014.

Underspecified landmark descriptions as in (c-d) above are likely to be felicitous in simpler, visual domains as those discussed in, e.g., [2]. Allowing target and landmark descriptions to disambiguate each other in these cases may arguably demand little cognitive effort from either speaker or hearer. In more complex situations, by contrast, underspecified landmark descriptions seem to be best avoided. For instance, 'the man behind the door' may demand considerable search effort if, e.g., there is a large number of potential landmarks (i.e., doors) to be inspected [13].

Leaving aside the issue of what kinds of domain may favour landmark underspecification, we will focus on simple situations of reference in which landmark underspecification is most likely frequent, and we will ask under which circumstances this may be preferred over full-specification. In addition to that, since the situations under consideration are simple, we will also investigate whether minimal descriptions are common – in line with studies such as [4] – or not [13].

The present investigation will be carried out as a controlled experiment to collect referring expressions under conditions that favour landmark underspecification. The resulting data set shows that underspecified landmark descriptions are not only common but, under certain conditions, may be largely preferred over minimally and fully-specified descriptions alike.

2 Related Work

Underspecification may be a natural way of referring to landmark objects. This seems to be the case at least in simple visual scenes as discussed in [2]. Accordingly, a number of REG algorithms focused on brevity or minimality may produced underspecified landmark descriptions, although this seems to be largely a side-effect of the main reference strategy, e.g., [4,2,10].

Finding empirical evidence of landmark reference underspecification is however difficult. Referring expression corpora are ubiquitous in NLP (e.g., for anaphora resolution as in [1]), but they lack the necessary semantic 'transparency' [7]. On the other hand, there are few publicly available, semantically annotated REG corpora conveying relational descriptions, but these generally do not seem to offer support to our current investigation.

One possible source of relational descriptions is the GIVE-2 corpus of instructions in virtual environments [6]. In a set of 992 definite descriptions of button object extracted from GIVE-2, 467 (47.1%) descriptions were found to use some kind of relation to a landmark object (e.g., doors, pieces of furniture etc.) However, landmark objects in GIVE-2 have few referable attributes, and most objects do not present variation in size, colour or shape. As a result, these objects are usually referred to by making use of the *type* attribute alone (e.g., 'the door'), with little variation in the way they may be (under or fully) specified.

Closer to our present interests, there is the case of the GRE3D and GRE3D7 corpora of referring expressions [5,17]. Both GRE3D and GRE3D7 are fully annotated and represent situations of reference in visual scenes in which the use of spatial relations was likely to occur (e.g., 'the cube next to the large sphere'). There are 224 relational

descriptions in GRE3D, and further 600 in GRE3D7. In GRE3D, the relation between target and landmark is always unique, whereas in GRE3D7 it is not.

We performed a brief analysis of the number of underspecified landmark descriptions in GRE3D and GRED7. Results are summarized in Table 1. As we shall focus on landmark underspecification, the top row – representing target underspecification – is presented for completeness only.

Table 1. Underspecified relational descriptions in the GRE3D and GRE3D7 data

Underspecification	GRE3D		GRE3D7		Overall	
target only	21	9.4%	3	0.5%	24	2.9%
landmark only	15	6.7%	36	6.0%	51	6.2%
target and landmark	22	9.8%	18	3.0%	40	4.9%
total	58	25.9%	57	9.5%	115	14.0%

From these results we observe that landmark underspecification (as would be produced by an algorithm such as [2]) is not infrequent. Over 15% of the descriptions in GRE3D show landmark underspecification, and even in the more complex scenes from GRE3D7 landmark underspecification is at 9%. This contrasts, for instance, studies such as [12,13], in which this kind of underspecification is shown to hinder identification in larger or structurally complex domain structures. In absolute numbers, however, both GRE3D and GRE3D7 corpora still lack sufficient evidence to support a study on landmark underspecification. For that reason, we decided to design a controlled experiment to collect referring expressions of this kind.

3 Current Work

3.1 Experiment Design

We designed a simple within-subjects experiment to investigate the generation of underspecified and minimal descriptions of landmark objects. The experiment makes use of near-identical pairs of visual scenes (kept as simple as possible to encourage minimal and/or underspecified descriptions without the identification risks discussed in [13]). Two examples are provided in Figure 1. The difference between the two is that on the left scene the landmark $q6$ has a uniquely distinguishing colour (white), whereas on the right side it does not.

In all scenes, a reference to the target object pointed by an arrow (e.g., $e2$) will most likely require a reference to the nearest landmark object (i.e., the box $q6$). Even when there are several similar landmark distractors (i.e., boxes) available, an unambiguous underspecified landmark description as in 'the star next to the box' is always possible, although with different degrees of difficulty.

From a REG perspective, the two scenes offer different choices in case the speaker decides to fully specify $q6$. Following the hierarchy of cognitive effort of spatial relations in [8], in the first scene the speaker may arguably prevent landmark underspecifica-

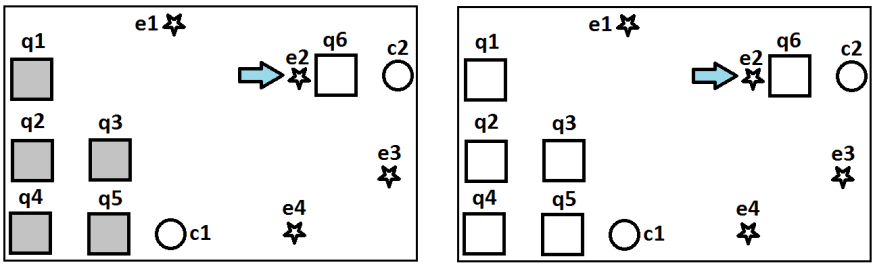

Fig. 1. Two situations of reference to a target (e2) via landmark object (q6).

tion simply by making use of an *absolute* reference form (e.g., colour).[2] This, according to [8], may require little cognitive effort. In the second case, by contrast, avoiding landmark underspecification may require a *projective* reference form (e.g., 'on the right'), or some less accessible attribute (e.g., ordinal reference as in 'the second box from left to right' etc.). This, also according to [8] may require more cognitive effort.

In our experiment setting, the left side of Figure 1 is an example of potential *absolute* (*abs*) reference, and the right side is an example of potential *projective* (*proj*) reference. The main goal of the experiment is to investigate in which of these two situations landmark underspecification (e.g., 'next to the box') is more common. In addition to that, given that we make use of simple visual scenes that are likely to encourage underspecification, we will also investigate whether minimal descriptions are common at all.

A landmark description is assumed to be underspecified only if conveying the single attribute *type* as in 'the box'. Any reference to colour, position or any other information that helps ruling out landmark distractors represent full-specification. Thus, typical underspecified landmark descriptions include 'the star *next to the box*' or 'the white star *next to the box*', since in our experiment setting the target colour is never uniquely distinguishing. By contrast, in 'the star *on the right side, next to the box*' the reference to the screen position, although unnecessary, helps landmark disambiguation, and it is therefore assumed to be fully-specified.[3] Minimal descriptions are always as in 'the star next to the box' or involving a similar spatial relation.

To gain further insight on these issues, both situations will be tested in three context sets of different sizes, conveying 1, 3 or 5 landmark distractors each. This gives rise to (2 situations * 3 context sizes =) 6 experiment conditions hereby called abs1, abs3, abs5, proj1, proj3 and proj5. For instance, the left side of Figure 1 represents the abs5 condition, and the right side represents the proj5 condition. Our two research hypotheses are as follows:

h1 Landmark underspecification is more frequent than full-specification when it is not possible to distinguish the landmark from other objects of the same type by means of an absolute reference form (e.g., colour).

[2] For a discussion on the semantics of the colour attribute see [18].

[3] Or, to be more precise, not underspecified.

h2 Minimal descriptions are overall less frequent than non-minimal descriptions across all conditions.

Hypothesis *h*1 will be tested by comparing the number of underspecified landmark references produced in all situations in which the landmark colour is unique (i.e., abs1, abs3 and abs5) with all situations in which landmark colour is not unique (i.e., proj1, proj3 and proj5). We would like to show that landmark underspecification is more frequent in the latter, that is, when a uniquely identifying colour is not available, and obtaining full-specification would presumably require more cognitive effort to produce and interpret. Conversely, we would like to show that landmark underspecification is less frequent when a uniquely distinguishing colour is available, presumably because full-specification is easier.

Hypothesis *h*2 will be tested by comparing the number of minimal and non-minimal descriptions in all situations and context sizes. We would like to show that minimal descriptions are less frequent that non-minimal descriptions even when the context set is quite simple as in our experiment setting.

3.2 Subjects

We recruited 73 Information Systems students who replied to an invitation made by email, and who agreed to volunteer. Participants were on average 20.9 years old and mostly male (62, or 84.9%). All participants were native speakers of Brazilian Portuguese.

3.3 Procedure

Participants were asked to run an executable (jar) file attached to the invitation e-mail and follow the instructions on screen. Upon execution, a brief instruction page was presented, informing the participant that she was about to volunteer anonymously for an experiment on Computer Science. A student id number was required as a means to collect gender and average age information.

Participants were informed that the task consisted of completing the sentence 'The object pointed by the blue arrow is the...' in a series of images, and that they should do so as naturally as possible, as if talking to a friend about the objects seen on screen, and free of ambiguity. In order to prevent biased answers, no examples were provided.

Each screen showed a different stimulus in random order, the command sentence and a text field accompanied by a 'Next' button. Simple cases of ambiguity as in 'the star' were automatically checked and, when necessary, an error message as in 'I do not know which star you are talking about. Please be more specific" was displayed. Other kinds of ambiguity were not automatically verified.

By providing an answer in the text field and pressing the 'Next' button, the next stimulus was displayed. At the end of the experiment, an encrypted file containing the participant's answers was produced, and the participant received instructions on how to e-mail her answer file back to the researchers in charge.

3.4 Materials

We used purpose-built software for presenting the experiment instructions and the stimuli (in random order), collecting the participant's answers and saving the data onto an encrypted file. The stimuli consisted of 11 images (6 representing our present research questions, and 5 fillers). The relatively large number of fillers was necessary to prevent answer patterns, as all 6 research questions could be in principle answered with the same (underspecified) description as in 'the star next to the box'.

The actual images used in the experiment are similar to the two scenes seen in Figure 1, but without object labels (presently added for ease of discussion). The target is always a star, and it is accompanied by three identical distractors, which forces subjects to add information for disambiguation. Describing the target will usually involve referring to the nearest landmark object, which is always a box.

3.5 Results

We collected 803 descriptions produced by 73 participants, who took on average 5.5 minutes each to complete the task. Since the task was performed without supervision and the participants did not receive examples of the expected description, the collected data set was manually verified for correctness. As a result, nine participants (12%) were identified as outliers. This was the case of participants who provided highly ambiguous descriptions, as in 'a white star', and who most likely misunderstood the task.

After removing the data produced by the outliers, our final data set contained 704 descriptions produced by 64 participants. For the purpose of the present study, a subset of 320 descriptions represents filler situations. These descriptions are not considered in the analysis to follow, which is solely based on the subset of 384 descriptions produced in the six situations of interest (abs and proj, with 1, 3 or 5 distractor landmarks each, cf. Section 3.1).

3.6 Analysis

We use χ^2 to compare description counts. The number of distractor landmarks (1, 3 or 5) had no significant effect on either $h1$ or $h2$. For that reason, in what follows we will consider mean frequencies for each condition group (abs 1/3/5 versus proj 1/3/5). Table 2 shows descriptive statistics for both hypotheses.

Table 2. Descriptive statistics for hypotheses h1 (left) and h2 (right).

	Underspec.		Fully-spec.			Minimal		Non-min.		
Cond.	mean	sdv	mean	sdv		mean	sdv	mean	sdv	
abs	11.7	1.5	50.3	3.8	31.0	7.3	0.6	55.0	2.6	31.2
proj	43.3	4.0	18.0	3.6	30.7	18.7	1.5	43.0	1.0	30.8
	27.5		34.2			13.0		49.0		

Most participants used both under- and fully-specified landmark descriptions. Only 8 participants (12.5%) always underspecified, and only 10 participants (15.6%) always fully-specified. According to Table 2 (left), landmark underspecification is less frequent than full specification when a uniquely distinguishing landmark colour is available (abs), and more frequent otherwise (proj). The difference is highly significant ($\chi^2 = 98.5, df = 1, p < 0.0001$). This confirms hypothesis $h1$.

Minimally distinguishing descriptions were overall rare. A total of 39 participants (60.9%) never produced minimal descriptions, and only 8 participants (6.3%) always produced them. As seen in Table 2 (right), minimally descriptions are less frequent than non-minimally distinguishing ones in both condition groups (abs and proj). The difference is highly significant ($\chi^2 = 18.12, df = 1, p < 0.0001$). This confirms hypothesis $h2$.

3.7 Further Issues

In addition to testing hypotheses $h1$ and $h2$, we performed a post-hoc analysis of description lengths and attribute usage across experiment conditions. Regarding description length, each description contained on average 3.1 attributes besides *type*, and there was no significant variation across experiment conditions. Since participants were free to complete each sentence with or without a noun (e.g., 'the star'), the use of *type* cannot be taken as indicative of a particular reference strategy, and is not presently analysed.

As for other attributes, there were only two significant difference across experiment conditions: first, the use of screen position attributes (e.g., 'the star *on the top-right corner*') increases three-fold when no absolute reference form is available (proj) ($\chi^2 = 17.09, df = 1, p < 0.0001$), and the use of landmark attributes (e.g., 'the star next to the *white* box') has a 28% increase when an absolute reference form is available (abs) ($\chi^2 = 11.35, df = 1, p < 0.0008$). Both results were to be expected as the absence of an absolute landmark attribute calls for an alternative (e.g., referring to the relative screen position) and, conversely, the presence of a discriminatory colour enables this reference strategy.

4 Conclusions

This paper described a controlled experiment to collect referring expressions under conditions that favour landmark underspecification. Results show that, at least in the simple visual scenes under consideration, the use of underspecified landmark descriptions is highly frequent, and that underspecification is preferred when full-specification seems more difficult to obtain (e.g., when the landmark object does not have a distinguishing colour).

As future work, we intend to apply some of these insights to the design of a more informed REG algorithm to produce definite descriptions conveying the appropriate level of information (e.g., under versus fully-specification). Preliminary results regarding this issue are presented in [16].

Acknowledgments. The authors acknowledge support by USP and FAPESP.

References

1. Cuevas, R.R.M., Paraboni, I.: A machine learning approach to portuguese pronoun resolution. LNAI 5290, 262–271 (2008)
2. Dale, R., Haddock, N.J.: Content determination in the generation of referring expressions. Computational Intelligence 7, 252–265 (1991)
3. Dale, R., Reiter, E.: Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive Science 19 (1995)
4. Dale, R.: Cooking up referring expressions. In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics. pp. 68–75 (2002)
5. Dale, R., Viethen, J.: Referring expression generation through attribute-based heuristics. In: Proceedings of ENLG-2009. pp. 58–65 (2009)
6. Gargett, A., Garoufi, K., Koller, A., Striegnitz, K.: The GIVE-2 corpus of giving instructions in virtual environments. In: Proceedings of LREC-2010 (2010)
7. Gatt, A., van der Sluis, I., van Deemter, K.: Evaluating algorithms for the generation of referring expressions using a balanced corpus. In: ENLG-07 (2007)
8. Kelleher, J.D., Costello, F.J.: Applying computational models of spatial prepositions to visually situated dialog. Computational Linguistics 35(2), 271–306 (2009)
9. Krahmer, E., van Deemter, K.: Computational generation of referring expressions: A survey. Computational Linguistics 38(1), 173–218 (2012)
10. de Lucena, D.J., Pereira, D.B., Paraboni, I.: From semantic properties to surface text: The generation of domain object descriptions. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial 14(45), 48–58 (2010)
11. de Novais, E.M., Paraboni, I.: Portuguese text generation using factored language models. Journal of the Brazilian Computer Society pp. 1–12 (2012)
12. Paraboni, I.: Generating references in hierarchical domains: the case of Document Deixis. Ph.D. thesis, University of Brighton (2003)
13. Paraboni, I., van Deemter, K.: Reference and the facilitation of search in spatial domains. Language and Cognitive Processes online (2013)
14. Pereira, D.B., Paraboni, I.: A language modelling tool for statistical NLP. In: Proceedings of TIL-2007. pp. 1679–1688 (2007)
15. Pereira, D.B., Paraboni, I.: Statistical surface realisation of Portuguese referring expressions. LNAI 5221, 383–392 (2008)
16. Teixeira, C.V.M., Paraboni, I., da Silva, A.S.R., Yamasaki, A.K.: Generating relational descriptions involving mutual disambiguation. LNCS 8403, 492–502 (2014)
17. Viethen, J., Dale, R.: GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In: Proceedings of UCNLG+Eval-2011. pp. 12–22 (2011)
18. Viethen, J., Goudbeek, M., Krahmer, E.: The impact of colour difference and colour codability on reference production. In: CogSci-2012. pp. 1084–1098 (2012)