A case study on context-bound referring expression generation

Maurice Langner
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum

Maurice.Langner@rub.de

Abstract

This paper describes and discusses the results of an empirical study on the production of referring expressions in visual fields with different object configurations of varying complexity and different contextual premises for using a referring expression. The visual fields are set up using data from the TUNA experiment with plain random or pragmatically enriched configurations which allow for target inference. Different categories of the situational contexts, in which the referring expressions are produced, provide different degrees of cooperativeness, so that generation quality and its relations to contextual user intention can be observed. The results of the study suggest that algorithms for REG must integrate individual generation preference and the cooperativeness of the situational task in order to model the broad variance between speakers more adequately.

1 Introduction

In the past, experiments on the production of referring expressions (REs) produced corpora on domains of different complexity, among those the TUNA corpus (van der Sluis, Gatt, van Deemter, 2006; 2006 online manual), GRE3D3 and GRE3D7 (Viethen & Dale, 2008;2011), ReferIT (Kazemzadeh et al., 2014), Wally (Clarke et al.,2013) and some interlingual experiments revealing that the basic concepts of reference are independent from language expertise (e.g. Khan & Siddiqui, 2015). Da Silva Rocha & Paraboni (2018) distinguish two general experimental designs in the REG task, related to the speaker-listener configuration: monologue and dialogue. The authors remark that "both dialogue and monologue are of course instances of real language use but, at least from these studies, it is not entirely clear whether the two situations are truly comparable" (p.2994). Questionable is still, whether or not content determination and the resulting generation quality, i.e. underspecification, minimality or overgeneration, may differ not only according to the speaker-listener configuration but also according to the context in which the REG task is situated. This question also includes variance between speakers. The experiment described in this paper builds on its predecessors, focusing on the technical and contextual parameters that may trigger differences in generation quality and content determination during production. The goal is to provide empirical data clarifying the influence of the situational context on the generation quality of referring expressions.

2 Methods

The experiment is designed using the TUNA furniture corpus and a subset of the TUNA people corpus that has been selected in a balanced way, making each feature value combination unique. It is conducted as a web-based experiment. Data from native speakers of English is collected using the crowdsourcing platform Amazon Mturk. The compiled corpus consists of 1029 production sessions from 50 participants.

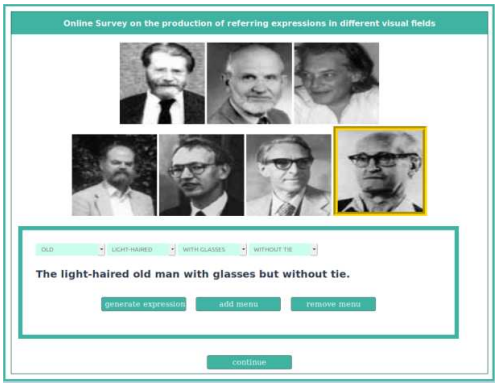

Figure 1: Web application: a production session

Proceedings of The 12th International Conference on Natural Language Generation, pages 19–23,
Tokyo, Japan, 28 Oct - 1 Nov, 2019.

category	context formulation (furniture)
+	You want to buy a very rare and valuable piece of furniture that you have been looking for for a long time. Please describe to the salesman which piece of furniture in his showroom you long for.
−	You are talking to friends about the design of your living room. They want to know which piece of furniture you recently sold on the internet. Please tell them.
o	You rearrange your living room. Tell your friend which piece of furniture you want to move to the free space below the windows.
category	**context formulation (people)**
+	You are the victim of a crime. Please describe to the police officer who of the suspects in the interrogation room is the criminal.
−	You want to buy a car and the sales agent wants to know who of his colleagues in the salesroom gave you advise at your previous visit. Please give a description.
o	You work as a waiter. You tell your colleague whom of the guests you still have to bring the bill. Please describe the guest.

Table 1: Situational contexts used in production sessions

domain/ID	furniture/A1IV1O890FN1QA-1
context	-
distractors	{ desk, front, grey, large}, { chair, left, green, small}, { fan, left, blue, large }
target	{sofa, left, green, small}
RE	*The sofa*
quality	minimality
domain/ID	furniture/AXQDSQGBC79S2-15
context	o
distractors	{fan, front, red, large}, { chair, left, red, large}, { sofa, front, green, large }
target	{ desk, back, blue, large}
RE	*The blue desk*
quality	overgeneration
domain/ID	people/A1TUI63TLCHYMR-12
context	-
distractors	{old, beard+, glasses-, hair-, front, shirt-, suit+, tie+}, {old, beard-, glasses-, hair+, front, shirt-, suit+, tie+}
target	{old, beard-, glasses-, hair+, front, shirt-, suit+,tie-}
RE	*the old front man with suit but without tie, shirt or glasses*
quality	overgeneration

Table 2: Examples from the experimental data

Production sessions were associated with different contexts which are representative of different communicative intents of the dialogue. Contexts are given in table 1. Either no contextual text was given and the participants were asked to generate expressions to their liking, or the context type was randomly chosen according to the domain type of the production session.

The contexts marked with + are designed with focus on the speaker's interest. In these contexts, the speaker envisages some personal intention for which it is important to convey to the listener which object he/she refers to. Correct identification is important to the speaker. The contexts marked with O are designed as rather neutral, where correct identification is of equal importance for both speaker and hearer in a collaborative task. The - marker indicates that these contexts focus on the hearer's interest, implied by the fact that the production task is the answer to the hearer's question. Correct identification is more important to the hearer than to the speaker.

3 Results

In this experiment, the main parameter of potential influence on the generation quality is the situational context. Consequently, all context conditions need to be evaluated in regard to overgeneration, minimality and underspecification. Examples of referring expressions produced by the participants and the corresponding context condition as well as the general session configuration are given in table 2.

The absence of a situational context (NONE condition) results in a nearly equal distribution of overgeneration and minimality (36.3% and 34.2%), while underspecification is slightly lower with a percentage of 29.5% (compare figure 2).

In contrast to this, the neutral context marked with O has a significantly higher ratio of minimal expressions, while overgeneration is close to equal in comparison. Underspecification occurs much less frequently (19.8%) than in sessions without situational context. The resulting difference between O and NONE is significant (χ^2 :5.66; $p < 0.05$). The + marked context shows a nearly equal distribution of overgeneration and minimality (32.0% and 31.0%), while there is a slight tendency towards underspecification (36.9%). The results for sessions with - marked contexts are diametrical to the + contexts (not significantly, though), revealing an approximately mirrored distribution of underspecification and minimality (32.4% and 30.0%), while overgeneration is slightly ahead with a ratio of 37.6%. Neither + nor - are significantly different from the sessions without context (NONE condition) but both are significantly different from the neutral O context (χ^2: $11.37/10.15$, $p : 0.003/0.006$).

The positive and negative contexts show tendencies towards overgeneration and underspecification respectively, but in opposite relation to the prior expectation. Contexts marked with +, in contradiction to intuitive assumptions, trigger more underspecification. A possible explanation for this is that the speaker may pay less attention to unique identification because it is only important to him-

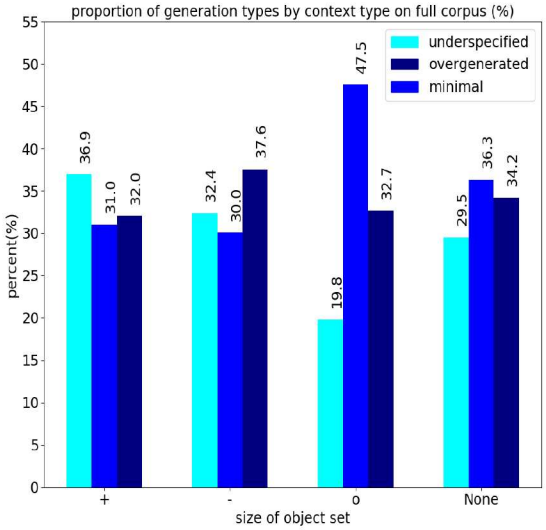

Figure 2: percentage of generation qualities by context condition

self/herself and not to the listener. In consideration of realistic settings of the REG task within dialogue speakers may be accustomed to the ability of correcting a referring expression after hearer feedback (installment noun phrases) in case that unique identification is not yet possible (Clark and Bangerter 2004, p.37). This would correlate with the tendency found by Paraboni & da Silva Rocha (2018) that dialogue settings trigger less overgeneration than monologue settings. Contexts marked with - provide a reversed balance of importance. The speaker may anticipate the listener's inability to infer the target, overgeneration being the consequence of guaranteeing the ability to correctly identify the target in consideration of its importance to the listener (Goodman & Frank, 2016). There is no reliable evidence for these assumptions due to data sparseness, but further research on larger data sets may reveal whether this relation is stochastically significant, resulting in a modelling of the REG task implementing context as pragmatic factors of cooperative dialogue.

For the neutral contexts, a cooperative task may result in a higher ratio of minimality for two reasons. Firstly, importance of correct identification prevails for both speaker and listener. Consequently, underspecification would be uncooperative, therefore being the disfavoured generation quality. This is mirrored by the low ratio of 19.8% for underspecified phrases. The second reason limiting the usage of overspecification may be not to give redundant information, which may cause the listener to reason about why the speaker violated Gricean Maximes of relevance, resulting in unnecessary cognitive effort prolonging the identification process or even in erroneous target inference (Paraboni et al., 2017). The data proves that situational contexts trigger significant differences in generation qualities. This depends on cooperativeness as much as on the quality and personal concern with the contexts, as well as the character of and relation to the listener.

3.1 Variance between speakers

For individual participants, the ratio of significantly different balances of generation qualities across context conditions is 8.16% (Fisher's exact test). This result for the variance within speakers is hardly reliable due to data sparseness [1]. Further studies with more sessions per participant will permit a valid evaluation of variance within speakers.

Nevertheless the experimental data gives rise to the assumption that variance between speakers is large. For each participant, sessions are counted according to the generation quality. The scatter plot reveals some interesting relations (see figure 3).

Every data point represents a single participant, its coordinates the counts of sessions where a minimal, overgenerated or underspecified expression has been produced. The data points are all arranged on a triangular surface. Each tip of the triangle represents a different group of participants with a specific preference to one generation type. The threshold used for the markers is a proportion of 0.6 of the preferred quality, 0.4 the combined occurrence of the residual qualities. As the plot already visualizes there is a huge variance between speakers. Comparing the group of participants preferring minimality (blue dots) with the group producing mainly overspecified phrases (dark blue triangles pointing upwards), the groups are highly significantly different (ANOVA, $f :> 100, p :< 5e^{-10}$). The groups preferring underspecification (light blue triangles pointing downwards) are even more significantly different from overgeneration and minimality with f-scores of 248.01 / 165.66 and a residual p-value of $4.19e^{-13}$ / $5.93e^{-12}$. The balanced group marked by black diamonds is

[1] With 20 sessions per participant and four context conditions, each context condition occurred five times on average. The values for generation qualities for each context condition are therefore too small for reliable significance tests.

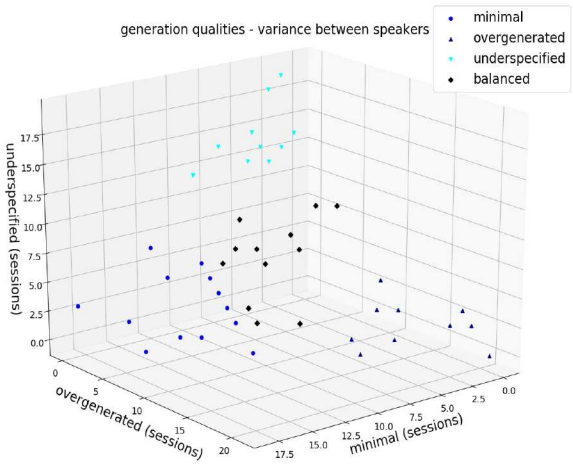

Figure 3: speaker variance

significantly different from all three groups with stronger preference towards a specific generation type, though f-scores range about 60 to 100 with p-values about $1e^{-8}$.

The graphs and the significant variance emphasize clearly that different speakers use different strategies in REG tasks (Viethen & Dale, 2011; Paraboni & Ferreira, 2014; van Deemter et al., 2010). The group preferring minimality is the largest with about 13 instances, overgeneration and underspecification each comprise about 10 participants. The least distinguished group shows a balanced use of generation qualities, representing a large amount of variance within speakers (van Deemter et al., 2010), since no clear tendency towards a strategy is visible. Further studies with more participants may probably allow for a Machine Learning approach (Janarthanam & Lemon, 2010, for further information on their Reinforcement Learning System for reference policies) in order to classify speakers according to instances of referring expressions produced during REG tasks. Another intriguing fact is that Goudbeek et al. (2011) were able to prove that the usage of dispreferred attributes as much as overgeneration can be primed in mixed comprehension and production sessions, showing that speakers may adjust even more to one another in real dialogue settings. For future work on reference in dialogue it is therefore crucial to control the parameter of strategical alignment and the mutual adjustment in cooperative reference tasks.

4 Discussion

The context condition triggers significant differences on the data gathered in this experiment. The neutral context entails a significantly higher ratio of minimality, proving that for the accomplishment of a cooperative task, participants tend to produce expressions by which the listener is able to identify the target unambiguously. The differences between the + and - contexts may be insignificant, but the existing tendency points in the opposite direction of the prior expectation. Participants produced slightly more overgenerated phrases in situations where target identification was more important to the listener, while underspecifying more in contexts where identification was more important for themselves. This indicates that speakers tend to value the listener's interests much higher than their own (van Deemter, 2016, p. 58). This needs to be considered since underspecification may be a symptom of habitual reference in realistic communication where correction and incremental reference is possible. Apart from the obvious fact that the experiment provided some evidence that contexts influence the REG task, the quality of formulations of exactly these contexts may not have been optimal. Further studies on more elaborated contexts integrating more factors of personal relationship towards the listener, dialogue settings, common ground and cooperativeness may show whether and to what degree the different parameters influence the content determination and the generation quality in context-bound REG tasks. Strategies are besides salience the most individual influence on the production of referring expressions. The variance between speakers clearly gave evidence for the existence of different generation strategies and preferences towards a specific generation quality. The distribution of instances of overgenerating and underspecifying instances are well balanced. Besides these there is a group of participants with balanced proportion of all three qualities giving empirical proof for the large variance within speakers. In a probabilistic approach to REG, the strategy of the listener and the speaker may be an important parameter the model has to integrate in order to adjust production and comprehension more elaborately to the relation between the interlocutors and their strategic alignment.

References

H. Rohde A. D. F. Clarke, M. Elsner. 2013. Wheres wally: the influence of visual salience on referring expression generation. *Frontiers in Psychology*, 4.

A. Gatt K. van Deemter, I. van der Sluis. 2006a. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Conference on Natural Language Generation, INLG-04*, pages 130–132, Sydney, Australia.

A. Gatt K. van Deemter, I. van der Sluis. 2006b. Manual for tuna corpus: Referring expressions in two domains.

A. Gatt K. van Deemter, I. van der Sluis. 2007. Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP-07*.

A. Gatt K. van Deemter, I. van der Sluis. 2010. Speaker- dependent variation in content selection for referring expression generation. In *Proceedings of the 8th Australasian Language Technology Workshop*, pages 81–89.

K. van Deemter. 2016. *Computational Models of Referring*. MIT Press, Cambridge, UK.

A. Bangerter H. H. Clark. 2004. Changing ideas about reference. In D. Sperber I.Noveck, editor, *Experimental Pragmatics*, pages 25–49. Springer.

M. Siddiqui I. Khan. 2015. Do speakers produce different referring expressions in their native language than a non-native language? *International Journal of Computational Linguistics Research*, 6(2):41–47.

D. da Silva Rocha I. Paraboni. 2018. Reference production in human-computer interaction: Issues for corpus-based referring expression generation. In *Proceedings of the 11th Language Resources and Evaluation Conference.*

M. de Sant'Ana F. Coutinho I. Paraboni, A. Lan. 2017. Effects of cognitive effort on the resolution of overspecified descriptions. *Computational Linguistics*, 43(2):451–459.

T. C. Ferreira I. Paraboni. 2014. Referring expression generation: Taking speakers' preferences into account. In I. Kopecek K. Pala P. Spjka, A. Horák, editor, *Text, Speech and Dialogue. 17th International Conference, TSD 2014*, pages 539–546. Springer International Publishing, Schweiz.

R. Dale J. Viethen. 2008. Generating relational references: What makes a difference? In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 160–168.

R. Dale J. Viethen. 2011. Gre3d7: A corpus of distinguishing descriptions for objects in visual scenes. In *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop*, pages 12–22.

M. C. Frank N. Goodman. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

O. Lemon S. Janarthanam. 2010. Adaptive referring expression generation in spoken dialogue systems: Evaluation with real users. In *Proceedings of Sigdial 2010: the 11th Annual Meeting of the special Interest Group on Discourse and Dialogue*, pages 124–131.

M. Matten T. Berg S. Kazemzadeh, V. Ordonez. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798.