Task demands and individual variation in referring expressions

Adriana Baltaretu and **Thiago Castro Ferreira**
Tilburg center for Cognition and Communication (TiCC)
Tilburg University
The Netherlands
{a.a.baltaretu,tcastrof}@tilburguniversity.edu

Abstract

Aiming to improve the human-likeness of natural language generation systems, this study investigates different sources of variation that might influence the production of referring expressions (REs), namely the effect of task demands and inter- intra- individual variation. We collected REs using a discrimination game and varied the instructions, telling speakers that they would get points for being fast, creative, clear, or no incentive would be mentioned. Our results show that task-demands affected REs production (number of words, number of attributes), and we observe a considerable amount of variation among the length of REs produced by single speakers, as well as among the REs of different speakers referring to the same targets.

1 Introduction

In Natural Language Generation, Referring Expression Generation (REG) is the task of generating references to discourse entities (Krahmer and Van Deemter, 2012). One of the most explored problems in REG is content selection, namely deciding what properties of the referent to include in a definite description, which is the focus of this work.

In general, REG algorithms have been developed on corpora collected with subtly different instructions. These nuanced instructions might have led to biases (e.g., influencing the types and frequency of attributes), which in turn could have led to biases in how REG algorithms operate, when trained on these corpora; or perhaps not. We propose a study investigating the effect of task demands on reference production. Moreover, REG typically focuses on generating unique descriptions by selecting content to distinguish the referent (target) from the other objects in the context (distractors). As result, computational models had been developed deterministically, always generating the same referring expression for a particular situation (Frank and Goodman, 2012; Van Gompel et al., 2012, for probabilistic models). This raises the question to what extent REs vary as a function of task demands and individual differences.

A number of studies have collected dedicated corpora of referring expressions, typically asking participants to produce distinguishable descriptions. However, these studies had nuanced instructions, and most of them relied on simple, schematic stimuli (grids of objects). For example, instructions emphasize accuracy and briefness (Viethen and Dale, 2010; van Deemter et al., 2006), introduce time pressure (Kazemzadeh et al., 2014) or use open ended formulations, asking participants to describe marked objects in such a way that they can be distinguished from other objects (Koolen et al., 2011). Task demands could influence the level of specification of the REs and the selection of (specific) attributes (Arts et al., 2011).

Another source of variation arises from speaker differences. Humans show individual style differences during language production, and speaker-dependent variation has been argued to be an important factor shaping the content of references (Viethen and Dale, 2010). Variation among individuals and across tasks has been proposed to arise from limitations of cognitive capacities of speakers and listeners (Hendriks, 2016). That individ-

Proceedings of The 9th International Natural Language Generation conference, pages 89–93,
Edinburgh, UK, September 5-8 2016.

ual variation exists is beyond doubt, however, we do not know of any studies to look at the amount of intra-individual variation (variation among the references of a same speaker) and inter-individual variation (variation among the references of different speakers in a same situation) in content selection using complex naturalistic scenes.

This paper focuses on human REs production in natural scenes, and we propose analysing whether RE production is influenced by task demands and speaker variation. We take a subset of stimuli and the instructions of an already existing reference game (Kazemzadeh et al., 2014) and ask participants to describe the object as best as possible (baseline condition), add time pressure, ask for creative and for clear REs. Compared to the baseline condition, we expect time pressure to trigger minimal short references with few adjectives; creativity to bring up novel and unusual ways of expressing attributes; clear REs to be longer and more detailed (more attributes). Regarding individual variation, we would like to measure to what extent REs of a speaker vary from each other, as well as the REs of different speakers for a same situation.

2 Methods

Participants Ninety native English speakers were paid to take part in the experiment via CrowdFlower, a crowdsourcing service similar to Amazon Mechanical Turk. We removed data from 17 respondents, as they declared not being native English speakers, not finishing, or misunderstanding the task. The final sample included 73 participants (31 males, mean age 38 years). The study followed APA guidelines for conducting experiments.

Materials Experimental materials consisted of 40 target objects, each presented in a different scene. These scenes have been semi-randomly selected from the larger set of images, illustrating aspects of everyday life, used to elicit REs in the ReferIt game (Kazemzadeh et al., 2014) . Our selection contains scenes that have at least one other object of the same type as the target, so as to elicit more than one word descriptions. To present participants with a wide range of objects, 20 scenes had animate targets and 20 scenes had inanimate ones. In each scene, the target was highlighted with a red bounding box, see

Figure 1: Experimental scenes depicting an animate target (above) and an inanimate one (below)

Figure 1.

Procedure Participants were randomly assigned to one of the conditions. We used and adapted the instructions of the ReferIt Game [1]. Participants' task was to produce distinguishable descriptions. For all conditions the instructions were identical except for the last sentence, that emphasized that participants should play fast (Fast condition, FA), be creative (creative condition, CR), clear and thorough (Clear and Thorough condition, CT) or no emphasis was be added (none condition, NO). Participants had to write down the description in a blank space provided under the scene. The scene remained on the screen until the participant introduced his description and pressed a button to continue. For each description participants received points, and were shown the score after submitting each description. The stimuli were presented in random order.

Analysis This study had a single independent variable Instruction type (levels: FA, CR, CT, NO) as between participants factor. The dependent variables were the length of the references (number of words), number of adjectives in a RE, type and frequency of adjectives (e.g., color, location), and number of unique words (words that occur only

[1] For the exact wording of the instructions see Annex 1

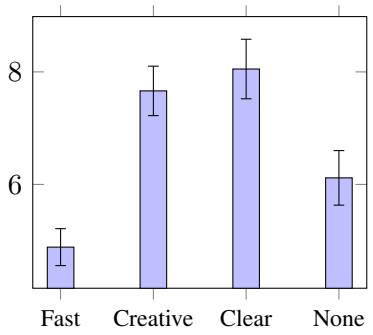

Figure 2: Average length of REs, split by condition. Error bars represent the 95% confidence intervals, y-axes represents mean number of words.

in a given condition). In order to test the observed differences, we conducted separate ANOVA tests. Individual variation was measured by computing the standard deviation of the REs' length. For intra-individual variation, standard deviation was measured for the group of REs produced by each speaker. For inter-variation, standard deviation was measured for the group of REs produced for each stimuli. Values close to zero indicate no intra- and inter- individual variation.

3 Results

In total 2920 references were produced (73 speakers * 40 scenes, FA, NO and CT conditions 18 participants each; CR condition 19 participants). The referring expressions consisted of a noun denoting the target object and all the phrases attached to it. Below, we report only significant effects.

Length of expressions There was a significant main effect of Instruction Type on the number of words, $F(3, 70) = 6.666$, $p = .01$, $\eta^2 = .222$ (see Figure 2). The FA condition had the shortest references ($M = 3.95$, $SE = .59$), followed by the NO condition ($M = 4.68$, $SE = .59$), the CT condition ($M = 6.36$, $SE = .59$) and the CR condition ($M = 7.19$, $SE = .56$). A post–hoc Tukey test showed that, compared to NO, only the CR and the CT conditions were significantly different ($p = .05$). The FA condition was significantly different from the CR ($p = .001$) and the CT conditions ($p = .006$).

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

Type	Examples	Frequency			
		Fast	Creative	Clear	None
location	man on the left	21%	32%	27 %	20%
color	white building	21%	30%	30%	19%
part	with balconies / with red nose	4%	37%	36%	23%
action	man holding a paper / bicycle being ridden	8%	35%	31%	26%
size	small monkey	18%	30%	34%	18%
emotion	smiling man	17%	33%	30%	20%
other		12%	35%	30%	23%

Number of adjectives There was a significant main effect of Instruction Type on the number of adjectives, $F(3, 70) = 4.362$, $p = .007$, $\eta^2 = .159$. The FA condition had the smallest number of adjectives ($M = .55$, $SE = .10$), followed by the NO condition ($M = .66$, $SE = .11$), the CT condition ($M = .88$, $SE = .10$) and the CR condition ($M = 1.03$, $SE = .99$). A post–hoc Tukey test showed that compared to the NO condition, there were no significant differences. The only significant difference was between the FA and the CR condition ($p = .008$) and there was an emerging trend suggesting a difference between the CR and the NO condition ($p = .065$).

Type and frequency of adjectives Speakers referred to the target objects using several types of attributes (see Table 1). In all conditions, the same types of attributes were present. The conditions with highest frequencies were CR and CL. The *other* category contains references with various attributes such as orientation (*dog facing left*), age (*the old building*), clothing (*the man wearing a white hat*), body descriptions (*the man holding his face in his hand*) and geographical origin (*the Indian man*).

Unique words Out of the total number of different words present in the corpus, the NO condition had 5% unique words, the FA condition 3% unique words, the CT condition 28% unique words and the CR condition 30% unique words.

Individual Variation Figure 3 depicts the intra- and inter- individual variation in the data. These results reveal, as one would expect, that there is indeed variation between participants in the amount of words they use for the same stimulus (($M = 6.09$,

$SD = 1.69$, $t(1, 39) = 22.406$, $p < 0.001$) and also variation between the stimuli (intra-individual variation, $M = 3.18$, $SD = 2.7$, $t(1, 72) = 14.80$, $p < 0.001$).

4 Conclusion and Discussion

Generally, content selection algorithms for definite descriptions generation behave deterministically by not taking into account factors like task demands or individual variation. The current paper investigated these two possible factors in the generation of definite descriptions, aiming to improve the *human-likeness* of NLG systems.

In particular, results showed that task demands (such as asking speakers to be fast, clear or creative) influences REs. Speakers who had to describe fast produced shorter references with less adjectives than the baseline condition. We assume that speakers in the fast condition may have lacked time and cognitive capacity to produce detailed references. Contrastively, speakers who had to be creative or clear produced longer and more detailed references. For example, the monkey in Figure 1 would be described as: *jumping monkey, FA*; *a primate showing off his business end, CR*; *small monkey with a very long tail, CT*; *a monkey on a persons' head, NO*. An interesting point for future research would be to investigate speaker's strategies across the four conditions, and to assess the accuracy with which listeners would be able to find the correct targets. Moreover, an open question remains how would the same REG algorithm perform when trained on datasets collected with different instructions.

Surprisingly, we did not observe any difference between the creative and clear references. Participants produced similar long and detailed references and the same types of attributes could be found in all conditions. Yet, the number of unique words for each of these conditions does hint there might be some other type of differences. Less creativity can also be due to the expectations workers have from MechanicalTurk tasks, which usually do not involve a 'creative' component. Participants might have interpreted our request for creativity as a request for explicit and detailed REs.

Our results also suggest a considerable amount of variation among the REs of a single speaker as well as among the REs of different speakers for a same situation. This result is in agreement with studies like (Viethen and Dale, 2010). An interesting observation for future research is that the level of intra-individual variation is lower than the level of inter-individual variation. As far as we know there are no computational REG models that take both inter- and intra- individual variation into account, and we wonder to what extent this could improve the *human-likeness* of the generated REs.

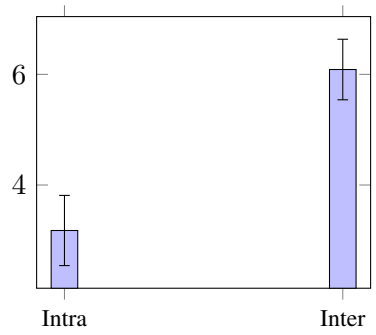

Figure 3: Average SDs of the length of REs per participant (intra-individual variation) and stimuli (inter-individual variation). Error bars represent 95% confidence intervals.

Annex 1. Instructions to the participants

Welcome to this game! In moments you will be shown a picture. In each picture there is an item bounded in red. Your goal is to describe the object as best as possible for another player, who has to select the object you describe. For each description you will earn points.

- *FA condition* The faster you play, the more points you win.
- *CT condition* The more clearly and thoroughly you describe, the more points you win.
- *CR condition* The more creative you are, the more points you win.
- *NO condition* Nothing

Acknowledgments

This work has been supported by the National Council of Scientific and Technological Development from Brazil (CNPq).

References

Anja Arts, Alfons Maes, Leo Noordman, and Carel Jansen. 2011. Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1):361–374.

Michael Frank and Noah Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Petra Hendriks. 2016. Cognitive modeling of individual variation in reference production and comprehension. *Frontiers in Psychology*, 7(506).

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the EMNLP*, pages 787–798, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ruud Koolen, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231–3250.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the Fourth INLG Conference*, pages 130–132. Association for Computational Linguistics.

Roger Van Gompel, Albert Gatt, Emiel Krahmer, and Kees Deemter. 2012. Pro: A computational model of referential overspecification. In *Proceedings of AMLAP*.

Jette Viethen and Robert Dale. 2010. Speaker-dependent variation in content selection for referring expression generation. In *Proceedings of the 8th Australasian Language Technology Workshop*.