Reference production in human-computer interaction: Issues for Corpus-based Referring Expression Generation

Danillo da Silva Rocha, Ivandré Paraboni
University of São Paulo, School of Arts, Sciences and Humanities
São Paulo, Brazil
rochadan60@gmail.com,ivandre@usp.br

Abstract

In the Natural Language Generation field, Referring Expression Generation (REG) studies often make use of experiments involving human subjects for the collection of corpora of definite descriptions. Experiments of this kind usually make use of web-based settings in which a single subject acts as a speaker with no particular addressee in mind (as a kind of monologue situation), or in which participant pairs are engaged in an actual dialogue. Both so-called monologue and dialogue settings are of course instances of real language use, but it is not entirely clear whether these situations are truly comparable or, to be more precise, whether REG studies may draw conclusions regarding attribute selection, referential overspecification and others regardless of the mode of communication. To shed light on this issue, in this work we developed a parallel, semantically annotated corpus of monologue and dialogue referring expressions, and carried out an experiment to compare instances produced in both modes of communication. Preliminary results suggest that human reference production may be indeed affected by the presence of a second (specific) human participant as the receiver of the communication in a number of ways, an observation that may be relevant for studies in REG and related fields.

Keywords: Natural Language Generation, Referring Expression Generation, Corpus

1. Introduction

In Natural Language Generation (NLG) studies, the collection of referring expressions - usually in the form of definite descriptions - produced by human subjects is a common task in Referring Expression generation (REG) and related fields (Krahmer and van Deemter, 2012). Descriptions of this kind are usually elicited from visual stimuli representing a context in which there is one particular target and additional distractor objects. Figure 1 illustrates a stimulus image from the Stars2 corpus (Paraboni et al., 2017a).

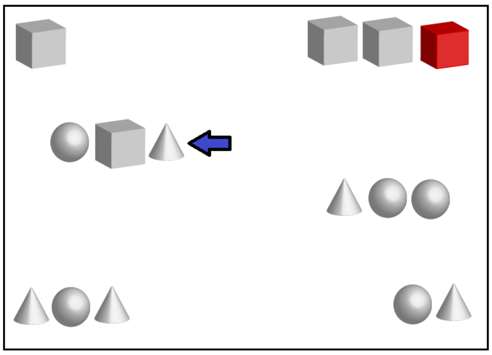

Figure 1: A stimulus image from the Stars2 corpus.

Given a context of this kind, the task of the human subject - who acts as a speaker or writer - is to produce a uniquely identifying description of the intended target. This could be accomplished, for instance, by producing a definite description as in 'The cone next to a grey box'.

Experiments involving human subjects for the collection of referring expression corpora are often implemented as a web-based data collection task, that is, without a particular addressee in mind. When there is no risk of confusion, we will hereby call these *monologue* situations[1].

So-called monologue situations are the method of choice for collecting data in TUNA (Gatt et al., 2007), GRE3D3 (Dale and Viethen, 2009), GRE3D7 (Viethen and Dale, 2011), Wally (Clarke et al., 2013) and other similar resources. By contrast, a number of data collection tasks make use of participant pairs in some form of dialogue. These include GIVE-2 (Gargett et al., 2010), ReferIt (Kazemzadeh et al., 2014), Stars2 (Paraboni et al., 2017a), *b5-ref* (Ramos et al., 2018) and others.

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 or, to be more precise, whether REG studies that rely on these methods may draw conclusions regarding attribute selection (Dale and Reiter, 1995), referential overspecification (Pechmann, 1989; Paraboni et al., 2017b) and others regardless of the mode of communication.

To shed light on this issue, in this work we developed a parallel, semantically annotated corpus of monologue and dialogue descriptions, and we present an experiment to compare definite descriptions produced in both modes of communication. The objective of the experiment is to investigate whether certain aspects of human reference production - which are particularly relevant for REG studies - may or may not be affected by the presence of a second human participant as the receiver of the communication.

The rest of this paper is organised as follows. Section 2 describes related work in the REG field. Section 3 presents our main experiment, whose results are discussed in Section 4 . Finally, Section 6 presents a number of conclusions and discusses future work.

[1] For a comprehensive discussion on this issue, we refer to (Ginzburg and Poesio, 2016).

2. Related work

This section briefly outlines existing work on REG methods and REG corpora.

2.1. Computational Referring Expression Generation

The attribute selection task for REG is generally modelled as an algorithm that receives as an input a context C comprising at least one target object r (or intended referent) and additional distractor objects. Objects are represented as sets of semantic properties, usually in the form of attribute-value pairs as in *colour*-red. The primary goal of a REG algorithm of this kind is to produce a uniquely identifying description L so as to distinguish r from every other distractor object within C.

Existing REG algorithms include early approaches such as the Greedy (Dale, 2002) and the Incremental (Dale and Reiter, 1995) algorithms. The Graph-based approach (Krahmer et al., 2003) allows the use of relational properties in a novel formulation of the task, and more recently the use of machine learning methods have been considered (Viethen and Dale, 2010; Ferreira and Paraboni, 2017). For a review of the main challenges in the field, see (Krahmer and van Deemter, 2012) and (van Deemter, 2016).

To illustrate the work of a typical REG algorithm, and to highlight a number of issues that may influence its outcome, let us consider a simplified visual context as in Figure 2.

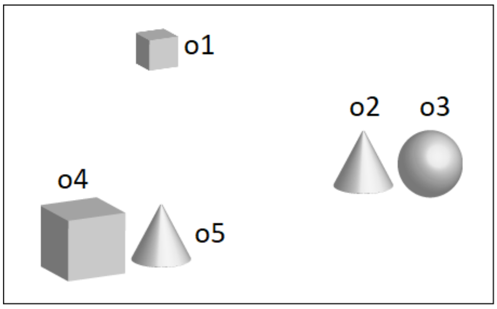

Figure 2: A visual context.

A scene of this kind may be represented as a knowledge base as follows.

o1 <*type*,box>,<*size*,small>

o2 <*type*,cone>,<*size*,large>,<*near*,o3>

o3 <*type*,ball>,<*size*,large>,<*near*,o2>

o4 <*type*,box>,<*size*,large>,<*near*,o5>

o5 <*type*,cone>,<*size*,large>,<*near*,o4>

Let us consider the goal of describing the target object r = $o5$ in this scene by making use of a domain-dependent list of preferred attributes P = <*type*, *size*, *near*>.

A standard attribute selection algorithm may start by making an empty set L (representing the output description) and then considering the first attribute in P, which in the present example is *type*. Since selecting *type* would rule out at least one distractor object (or, in this case, all objects that are not cones $o1$, $o3$ and $o4$), this attribute is included in the output description L, and the relevant objects are removed from the context C.

Next, the second attribute in P is considered, that is, *size*. Since all remaining objects in C share the same *type* value (large), the selection of *size* is disregarded.

Finally, the *near* attribute is examined. Since the target is the only object near $o4$, *near* is selected for inclusion in L and the context is emptied. The algorithm may now terminate or may be called recursively to describe $o4$ as well, resulting in a description that could be subsequently realised as in, e.g., ‘the cone near a/the large box’.

Attribute selection is usually driven by discriminatory power (Olson, 1970) (e.g., in the above example, only properties that rule out at least one distractor object are selected), and it is heavily influenced by the order P in which attributes are considered for inclusion in L. As a result, output descriptions may vary considerable both in length and in the kinds of information that they convey (Paraboni, 2003). Given that the ultimate goal of computational REG is (arguably) the generation of descriptions that resemble those produced by human speakers, corpora of human-produced referring expressions are often collected to gain insights on these issues, and in some cases to provide training data for machine learning REG models.

2.2. Corpora for REG

Data collection for REG (e.g., as training data for corpus-based approaches as in (Ferreira and Paraboni, 2014a; Ferreira and Paraboni, 2014b)) may in some cases lead to the development of a so-called REG corpora. In this section we briefly review some of the resources that are more directly relevant to standard REG, in the sense proposed in (Dale and Reiter, 1995) and others.

The TUNA corpus (Gatt et al., 2007) was implemented as a web-based data collection task involving single participants acting as speakers. The corpus comprises two domains: Furniture, containing descriptions of pieces of furniture (e.g., desks, chairs etc.), and People, containing descriptions of human photographs. Both Furniture and People scenes contain from three to seven objects each. The corpus was developed for the study of the content selection task of atomic descriptions, and as a dataset for a series of Shared Tasks (Gatt et al., 2009). TUNA contains 2280 descriptions produced by 60 speakers, being 1200 Furniture descriptions and 1080 People descriptions in so-called monologue situations.

The GRE3D3 and GRE3D7 corpora (Dale and Viethen, 2009; Viethen and Dale, 2011) were also implemented as web-based collection tasks involving single participants in the role of speakers. In both cases, the domain consisted of visual scenes containing either three (in GRE3D3) or seven (GRE3D7) geometric objects (boxes and balls) each, with limited variation in colour and size. The goal of the data collection was to investigate the use of *relational descriptions* (Krahmer et al., 2003; dos Santos Silva and Paraboni, 2015) as in ‘the ball next to the yellow cube’ in a context in which an atomic description (e.g., ‘the ball’) would suffice for the purpose of identification. GRE3D3 contains 630 descriptions produced by 63 participants, and GRE3D7 contains 4480 descriptions produced by 287 participants, in both cases presented in monologue situations.

The Stars2 corpus (Paraboni et al., 2017a) was implemented as a series of collaborative tasks involving speaker-hearer participant pairs. The domain consisted of visual scenes containing 15 objects each. The corpus was developed for the study of referential overspecification of atomic and relational descriptions alike. Stars2 contains 884 descriptions produced by 56 speakers in dialogue situations.

More recently, a number of data collection tasks for REG have relied on crowd sourcing methods. These include, for instance, the issue of reference in open domains such as the Referit corpus (Kazemzadeh et al., 2014), conveying descriptions of visual elements in real photographs. The tasks in this case involves describing vague target objects (e.g., the central region of a picture, which may be variously described as 'the old man', 'the middle of the picture', 'a person's nose' etc.), which may be considered distinguished from standard REG.

3. Current work

Unsupervised data collection for REG (e.g., as elicited in a web-based monologue situation) may arguably produce lower quality definite descriptions in general. For instance, without a particular hearer in mind, participants of a REG experiment may be less inclined to craft their referring expressions so as to be unambiguously understood. Moreover, even assuming that all speakers are sufficiently careful during the experiment, situations of communication that are deemed less critical or somewhat less important may still affect attribute choice and referential overspecification (Arts et al., 2011), issues that are at the heart of many REG studies. As a result, monologue and dialogue situations of communication may not be entirely comparable or, at the very least, may elicit different referring expressions.

To clarify this, we designed an experiment to compare definite descriptions produced in dialogue and monologue situations of communication. The goal of the experiment is to investigate whether certain aspects of human reference production - which are particularly relevant for studies in Referring Expression Generation - may or may not be affected by the presence of a second human participant acting as a hearer or reader.

The experiment consists of reproducing a number of trials presented in the Stars2 data collection task (which was originally carried out as a limited form of dialogue between speaker-hearer participant pairs) and by replacing the hearer participant for a simple web interface with no feedback. In other words, we intend to reproduce a number of Stars2 dialogues in monologue situations not unlike those implemented in a number of prominent data collection tasks for REG, including TUNA (Gatt et al., 2007), GRE3D3 (Dale and Viethen, 2009) and others.

3.1. Hypotheses

The experiment investigates four research questions concerning possible differences between definite descriptions produced in monologue and dialogue situations. These questions address both quantitative and qualitative aspects of reference production, and are based on the semantic annotation associated to the descriptions under evaluation as defined by the annotation scheme of the Stars2 corpus (Paraboni et al., 2017a). Generally speaking, our four research questions assume that, when a particular hearer is present (that is, in a true dialogue situation), speakers will design their referring expressions more carefully than if they were alone (that is, in a so-called monologue situation). This general principle is consistent with studies on referring expression generation in critical situations of communication (Arts et al., 2011) and others. The four hypotheses to be investigated are detailed as follows.

> *h1: Descriptions produced in dialogue situations are, on average, longer than those produced in monologue situations.*

Hypothesis $h1$ will be tested by comparing the average number of annotated properties produced in dialogue situations with descriptions produced under identical circumstances (i.e., in the same context) in monologue situations. Dialogue descriptions are expected to convey more information than monologue descriptions.

> *h2: Descriptions produced in dialogue situations contain, on average, more spatial relations than those produced in monologue situations.*

Hypothesis $h2$ will be tested by comparing the average number of spatial relations observed in dialogue and monologue situations. Dialogue situations are expected to present more relational descriptions.

> *h3: Atomic descriptions produced in dialogue situations include, on average, more information about the target object than those produced in monologue situations.*

Hypothesis $h3$ will be tested by measuring the level of target overspecification in atomic descriptions (e.g., 'the box' versus 'the *red* box' in a context in which there is only one box object) produced in dialogue and monologue situations alike. To this end, the level of overspecification is defined as the number of target properties beyond what would be strictly required for disambiguation, that is, the number of properties added to an otherwise minimally distinguishing description. Atomic descriptions in dialogue situations are expected to be more overspecified than those in monologue situations.

> *h4: Relational descriptions produced in dialogue situations include, on average, more information about the landmark object than those produced in monologue situations.*

Hypothesis $h4$ is similar to $h3$, but now focusing on the landmark portion of relational descriptions (e.g., 'the cube next to a *red* box' in a context in which there is only one box object). This hypothesis will be tested by measuring the level of landmark overspecification in dialogue and monologue situations alike. Descriptions of landmark objects in dialogue situations are expected to be more overspecified than those in monologue situations.

3.2. Subjects

24 native speakers of Brazilian Portuguese, on average 31 years-old and predominantly male (22), drawn from the same population as in (Paraboni et al., 2017a). All participants had normal or corrected vision.

3.3. Materials

24 trials[2] originally presented in the Stars2 data collection task. Each trial consists of a sequence of 16 stimulus images in the same order in which they were presented in the original experiment. The selected trials were those in which there was no misunderstanding or repetition, that is, those in which the hearer always managed to identify the target described by the speaker without any further clarification.

3.4. Procedure

The experiment followed the same procedure as in (Paraboni et al., 2017a), except that there was no hearer participant available, that is, each speaker worked on his/her own by interacting with a WEB interface without any feedback. Participants were required to provide basic information regarding age, gender and an informed consent. This was followed by instructions on how to complete the task. The instructions were the same as in the original experiment - essentially, participants were requested to uniquely identify the object pointed by an arrow - and they did not include any actual examples of referring expression.

After reading the instructions, participants were directed to the experiment proper. This consisted of presenting a series of stimulus images taken from (Paraboni et al., 2017a), one by one, and by requesting the participant to provide a description for the target of each scene. An initial set of four images was presented for practice only, and their responses were not recorded. The following 16 images represented the actual stimuli for the data collection task. During the practice session participants were allowed to ask questions regarding the task to the research assistant. After practice, participants were left unattended and they could not make further questions so as to establish the intended monologue communication setting.

4. Results

A set of 384 descriptions was collected and subsequently annotated by two judges following the same 19-attribute annotation scheme in (Paraboni et al., 2017a), plus an additional 'others' attribute intended to represent any other kind of information outside the scope of the original study. Put together, monologue and dialogue descriptions comprise a parallel corpus of semantically-annotated referring expressions to be made available for further studies on both modes of reference production.

Before discussing our research hypotheses, a preliminary test was carried out so as to assess the closeness between the two (dialogue and monologue) data sets by measuring Dice coefficients (Dice, 1945). As a result, an average Dice score of 0.72 was obtained, which may suggest that the difference between dialogue and monologue datasets was considerably high. The following analysis will discuss this difference in more detail.

Table 1 summarizes our main results for descriptions obtained both in dialogue and monologue situations, in which significant differences between dialogue and monologue descriptions are highlighted. Results are represented as the average description *length* (measured as the number of annotated properties according to the original annotation scheme in (Paraboni et al., 2017a)), the average number of spatial relations (*rel-count*), the average number of overspecified properties in the target portion of the descriptions (*over-tg*), and the average number of overspecified properties in the landmark portion of the descriptions (*over-lm*) as required for evaluating hypotheses $h1..h4$.

The number of description (n) considered in each case is either 384 (i.e., the entire dataset) or 192. The latter corresponds to the tests based on *over-tg* and *over-lm*, which focus on the half of the situations in which the use of relational properties was either optional or compulsory for disambiguation.

5. Discussion

Overall results in Table 1 in principle suggest that descriptions produced in dialogue situations convey, on average, more information than monologue descriptions. To verify this, a between-subjects ANOVA test was carried out for each variable (*length*, *rel-count*, *over-tg* and *over-lm*) as follows.

Regarding hypothesis $h1$, we notice that dialogue descriptions are, on average, longer than those produced in monologue situations. The difference is significant (F(1,101)=8.52, MSE=26.285 $p<0.05$). This offers support to hypothesis $h1$.

Regarding $h2$, we notice that dialogue descriptions are more likely to include a spatial relation than monologue descriptions. The difference is significant (F(1,101)=5.15, MSE=3.245 $p<0.05$). This supports $h2$.

Regarding $h3$, results suggest that dialogue target descriptions may be less overspecified than monologue descriptions. However, the difference was not significant. This outcome does not offer support to $h3$.

Finally, regarding $h4$, we notice that dialogue landmark descriptions are more overspecified than monologue descriptions. The difference is also significant (F(1,101)=4.635, MSE=1.622 $p<0.05$). This supports $h4$.

6. Final remarks

This paper presented an experiment to compare descriptions produced in dialogue and monologue situations, and provided a parallel corpus of semantically-annotated referring expressions in the two modes of communication. Despite the small scale of our study, preliminary results suggest a number of quantitative and qualitative differences between the two kinds of data collection, and which may arguably impact studies on content selection, referential overspecification and other issues that play a central role in REG and related fields.

As future work, we intend to compare monologue and dialogue descriptions against those produced in simulated dia-

[2] We reproduced Stars2 trials 52, 89, 105, 115, 136, 307, 439, 455, 503, 538, 585, 597, 621, 704, 788, 823, 832, 858, 895, 898, 969, 972, 978 and 998 from (Paraboni et al., 2017a).

Table 1: Dialogue vs. Monologue results

	h1: length		h2: rel-count		h3: over-tg		h4: over-lm	
	dial.	mono.	dial.	mono.	dial.	mono.	dial.	mono.
mean	**3.35**	2.98	**0.95**	0.82	0.59	0.51	**0.84**	0.71
var	3.57	2.60	0.68	0.58	0.47	0.32	0.38	0.32
n	384	384	384	384	192	192	192	192

logue, that is, by making use of purpose-made tool to play the role of the hearer participant in a traditional dialogue setting. In doing so, our long-term goal is to obtain referring expressions that resemble those that would be obtained from a real dialogue task at a lower cost, that is, by making use of fewer experiment participants.

The monologue-dialogue parallel corpus - hereby named Stars2MD - is available for research purposes upon request.

7. Acknowledgements

This work has been supported by grant # 2016/14223-0, São Paulo Research Foundation (FAPESP).

8. Bibliographical References

Arts, A., Maes, A., Noordman, L. G. M., and Jansen, C. (2011). Overspecification in written instruction. *Journal of Pragmatics*, 43.

Clarke, A. D. F., Elsner, M., and Rohde, H. (2013). Where's Wally: The influence of visual salience on referring expression generation. *Frontiers in Psychology*, 4(329).

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

Dale, R. and Viethen, J. (2009). Referring expression generation through attribute-based heuristics. In *Proceedings of ENLG-2009*, pages 58–65.

Dale, R. (2002). Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 68–75.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

dos Santos Silva, D. and Paraboni, I. (2015). Generating spatial referring expressions in interactive 3D worlds. *Spatial Cognition & Computation*, 15(03):186–225.

Ferreira, T. C. and Paraboni, I. (2014a). Classification-based referring expression generation. *Lecture Notes in Computer Science*, 8403:481–491.

Ferreira, T. C. and Paraboni, I. (2014b). Referring expression generation: taking speakers' preferences into account. *Lecture Notes in Artificial Intelligence*, 8655:539–546.

Ferreira, T. C. and Paraboni, I. (2017). Generating natural language descriptions using speaker-dependent information. *Natural Language Engineering*, pages 1–22.

Gargett, A., Garoufi, K., Koller, A., and Striegnitz, K. (2010). The GIVE-2 corpus of giving instructions in virtual environments. In *Proceedings of LREC-2010*.

Gatt, A., van der Sluis, I., and van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proc. of ENLG-07*.

Gatt, A., Belz, A., and Kow, E. (2009). The TUNA challenge 2009: Overview and evaluation results. In *Proceedings of the 12nd European Workshop on Natural Language Generation*, pages 174–182.

Ginzburg, J. and Poesio, M. (2016). Grammar is a system that characterizes talk in interaction. *Frontiers in Psychology*, 7(1938).

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. (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. Association for Computational Linguistics.

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

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

Olson, D. R. (1970). Language and thought: aspects of a cognitive theory of semantics. *Psychological Review*, 77(4):257–273.

Paraboni, I., Galindo, M., and Iacovelli, D. (2017a). Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*, 51(2):439–462.

Paraboni, I., Lan, A. G. J., de Sant'Ana, M. M., and Coutinho, F. L. (2017b). Effects of cognitive effort on the resolution of overspecified descriptions. *Computational Linguistics*, 43(2).

Paraboni, I. (2003). *Generating references in hierarchical domains: the case of Document Deixis*. Ph.D. thesis, University of Brighton.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27(1):98–110.

Ramos, R. M. S., Neto, G. B. S., Silva, B. B. C., Monteiro, D. S., Paraboni, I., and Dias, R. F. S. (2018). Building a corpus for personality-dependent natural language understanding and generation. In *11th International Conference on Language Resources and Evaluation (LREC-2018) (to appear)*, Miyasaki, Japan. ELRA.

van Deemter, K. (2016). *Computational Models of Referring: A Study in Cognitive Science*. MIT Press.

Viethen, J. and Dale, R. (2010). Speaker-dependent variation in content selection for referring expression generation. In *Australasian Language Technology Association ws.*, pages 81–89.

Viethen, J. and Dale, R. (2011). GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In *Proceedings of UCNLG+Eval-2011*, pages 12–22.