Referring Expression Generation in time-constrained communication

André Costa Mariotti, Ivandré Paraboni
University of São Paulo, School of Arts, Sciences and Humanities
São Paulo, Brazil
drehmariotti@gmail.com, ivandre@usp.br

Abstract

In game-like applications and many others, an underlying Natural Language Generation system may have to express urgency or other dynamic aspects of a fast-evolving situation as text, which may be considerably different from text produced under so-called 'normal' circumstances (e.g., without time constrains). As a means to shed light on possible differences of this kind, this paper addresses the computational generation of natural language text in time-constrained communication by presenting two experiments that use the attribute selection task of definite descriptions (or Referring Expression Generation - REG) as a working example. In the first experiment, we describe a psycholinguistic study in which human participants are engaged in a time-constrained reference production task. This results in a corpus of time-constrained descriptions to be compared with 'normal' descriptions available from an existing (i.e., with no time constraint) REG corpus. In the second experiment, we discuss how a REG algorithm may be customised so as to produce time-constrained descriptions that resemble those produced by human speakers in similar situations. The proposed algorithm is then evaluated against the time-constrained descriptions produced by the human subjects in the first experiment, and it is shown to outperform standard approaches to REG in these conditions.

Keywords: Natural Language Generation, Referring Expression Generation, time constraints

1. Introduction

In Natural Language Generation (NLG), the subtask of Referring Expression Generation (REG) (Krahmer and van Deemter, 2012) consists of providing linguistic forms to describe (or refer to) discourse objects. REG involves two distinct issues: content determination (or deciding what to say) and surface realisation (or deciding how to say it in a given language). In this work we focus on the former, addressing the issue of attribute selection of definite descriptions hereby called REG for simplicity.

Of particular interest to the present discussion is the task of generating natural language text from a visual context provided by, e.g., a video game application. One example of application of this kind is illustrated in Figure 1[1].

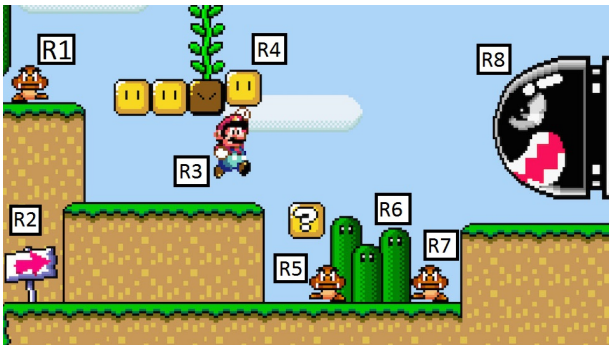

Figure 1: Example of a referential context

In this case, we may assume that an underlying NLG system would receive as an input the current state of the game, including all relevant objects in the scene and their properties, and would generate context-sensitive messages such as 'Try to jump as high as possible to avoid X!' in which X stands for a description of a particular target object r. For instance, assuming r to correspond to the object labelled as $R5$, the description X may be realised as 'the little monster under the question mark', 'the brown monster at the bottom', 'the nasty little thing in front of you', among many others.

The choice of contents to be expressed as a definite description is of course determined by its communicative goals (Paraboni and van Deemter, 1999). In certain game-like applications, however, we notice that these goals may be influenced by the need to express urgency or other dynamic aspects of a fast-evolving situation and, as a result, descriptions to be produced under 'normal' circumstances may be different from those that should be produced under time constrains or other atypical situations (Arts et al., 2011). Possible differences of this kind, and how time-constrained descriptions may be generated using existing REG algorithms, are the focus of the present work.

This paper addresses the computational generation of natural language in time-constrained communication by presenting two experiments that use the attribute selection task of definite descriptions as a working example. In the first experiment, we present a psycholinguistic study in which human participants are engaged in a time-constrained reference production task. From this study, a corpus of time-constrained descriptions - hereby called Stars2T - is elicited, and then compared with 'regular' descriptions available from an existing REG corpus.

Based on these initial findings, in the second experiment we discuss how a REG algorithm may be customised to produce time-constrained descriptions that resemble those produced by human speakers in similar situations by adding a certain amount of overspecified information. The proposed method is evaluated on time-constrained data, and it is shown to outperform standard approaches to REG in these (admittedly unusual) conditions.

The paper is structured as follows. Section 2 briefly de-

[1] Image captured from Nintendo's Super Mario World.

scribes previous work on REG and related topics. Section 3 describes the experiment evolving human subjects in a reference production task. Section 4 presents and evaluates our REG algorithm. Finally, Section 5 draws a number of conclusions and suggests future work.

2. Background

2.1. REG attribute selection

In the Natural Language Generation field, the attribute selection task for REG has been the focus of a wide range of computational methods (Dale and Reiter, 1995; Krahmer et al., 2003; Ferreira and Paraboni, 2014a), reference corpora (Gatt et al., 2007) and shared tasks (Gatt et al., 2009).

REG is typically implemented as an algorithm that receives as an input a context C containing a target object r and additional distractor objects. Objects are represented as sets of semantic properties, usually in the form of attribute-value pairs as in *type*-monster. For instance, the following is a possible representation of the context conveying the eight objects labelled R1..R8 in the previous Figure 1.

R1 <*type*,monster>,<*colour*,brown>

R2 <*type*,sign>,<*colour*,white>

R3 <*type*,man>,<*colour*,red>, <*size*,small>,<*below*,R4>

R4 <*type*,block>,<*colour*,yellow>,<*above*,R3>

R5 <*type*,monster>,<*colour*,brown>,<*left*,R6>

R6 <*type*,cactus>,<*colour*,green>,<*left*,R7>,<*right*,R5>

R7 <*type*,monster>,<*colour*,brown>,<*right*,R6>

R8 <*type*,bomb>,<*colour*,black>,<*size*,large>

Key aspects of the REG attribution selection task are easily illustrated by the Dale & Reiter Incremental algorithm (Dale and Reiter, 1995). The algorithm takes as an input an intended target object r, a context C containing a number of distractor objects from which r has to be distinguished, and a domain-dependent list of preferred attributes P.

The goal of the algorithm is to produce a set L of properties of r such that L distinguishes r from every other object in C. The original Incremental algorithm focuses on the selection of atomic properties only, but handling relational properties (e.g., 'left') is sufficiently straightforward as well (Krahmer et al., 2003; dos Santos Silva and Paraboni, 2015). In what follows we discuss an example that includes atomic and relational properties alike.

The Incremental algorithm iterates over the preferred order P and considers one attribute at a time. If an attribute a helps disambiguate the intended referent, a is included in the output description L. The algorithm terminates when L denotes r and no other object in C, or when all possible attributes in P have been considered. In the former case, L may be realised as a definite description as in 'the large bomb', and in the latter as an indefinite description as in 'a brown monster'.

To illustrate this, let us consider, for instance, the goal of describing the target r = $R5$ in the above context based on a preference order P = <*type*, *colour*, *left*>. The algorithm starts by making an empty set L and then considers the first attribute a in P, that is, *type*. Since this attribute rules out several distractor objects (namely, all objects that are not monsters, that is, $R2$, $R3$, $R4$, $R6$ and $R8$), a is included in L and the objects that have been ruled out by the operation are removed from C.

At this point, the context contains only two distractor objects left: $R1$ and $R7$, both of which being of the same type as r itself. Next in the preference order P, the *colour* attribute is considered. Since all objects in C share the same *colour* value (brown), this attribute does not rule out any distractor objects in C, and it is therefore disregarded.

Finally, the *left* attribute is considered. Since neither $R1$ or $R7$ are on the left side of anything sufficiently close for the purpose of reference (and therefore do not have a *left* property defined in the knowledge base), (*left-R6*) is also added to L, and both distractors are removed from C.

As a result, the context C is now empty, and the algorithm terminates by returning the expression L = {(*type*, *monster*), (*left*, *R6*)}, in which case $R6$ itself could be described recursively as {(*type*, *cactus*)}. This output description could be subsequently realised as, e.g., 'the monster on the left side of the cactus'. Similarly, $R1$ could have been (ambiguously) described as, e.g., 'a monster', and $R7$ as, e.g., 'the monster on the right side of the cactus'.

In addition to the fundamental goal of producing unambiguous descriptions, different REG algorithms may also consider additional requirements. In particular, a great deal of attention has been devoted to the question of *humanlikeness* or plausibility of the generated output (Gatt et al., 2009). As in many other NLG tasks, it is generally assumed that the output description should be as similar as possible to what human speakers would have produced under the same circumstances. This may involve the generation of brief descriptions (Dale, 2002), modelling domain preferences (Gatt et al., 2013), models of human variation (Ferreira and Paraboni, 2014b) and many others. For a detailed account of computational REG, we refer to (van Deemter, 2016).

2.2. Reference production in critical situations

The literature on time-constrained reference production is, from both linguistic and computational perspectives, scarce. The issue is however related to reference production in critical situations of communication, which is the focus of the study in (Arts et al., 2011).

The critical conditions considered in (Arts et al., 2011) are defined as situations in which the identification of a target object is understood by the participants of the dialogue as highly important for the success of an underlying task. To investigate the production of referring expressions in these situations, the study made use of an experiment involving human subjects engaged in a fictitious task of assisting a high-risk surgery that was occurring remotely, in which case the participants were requested to produce descriptions with minimal risk of misinterpretation.

In the experiment in (Arts et al., 2011) there were two groups of participants. The first group was in charge of producing descriptions under so-called 'normal' conditions, and the second group was in charge of producing descriptions under critical condition. Results of the comparison between the two (critical and non-critical) groups suggested that, among other findings, criticality affects the production of referring expressions. More specifically, the study

shows that criticality leads to *referential overspecification* (Pechmann, 1989), that is, the production of expressions containing more information than strictly necessary for disambiguation. An example of overspecified description for the target $R1$ in Figure 1 would be 'the *brown* little monster near the top-left corner', in which the reference to colour is not strictly necessary for identification, that is, it is said to be overspecified.

2.3. The Stars2 corpus

An investigation on the possible relation between referential overspecification and criticality (presently modelled as a time constraint) requires, at the very least, a linguistic domain in which referential overspecification is ubiquitous[2].
In this section we briefly describe the Stars2 corpus of referring expressions (Paraboni et al., 2017a) that will be the basis of the first experiment described in the next section.
Stars2 is a corpus of definite descriptions produced by human subjects in a controlled experiment involving situations of reference in which the use of relational properties between target and landmark objects (e.g., 'the cone next to a box') is likely to occur.
An example of referential context in this domain is illustrated in Figure 2.

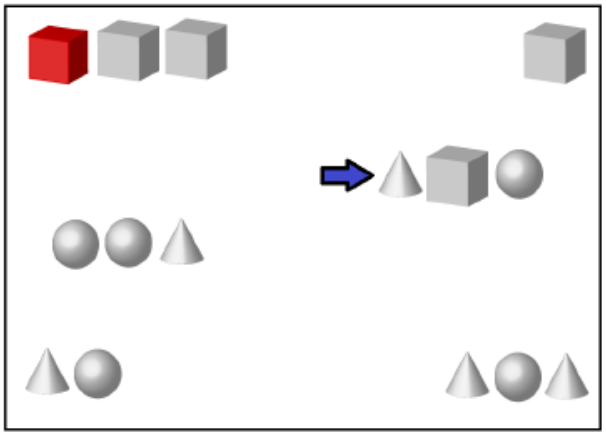

Figure 2: A context in Stars2

Based on stimulus images of this kind, participants of the data collection task were requested to produce a description that would allow the target object (in the example, the cone pointed by the arrow) to be uniquely identified. The goal in (Paraboni et al., 2017a) was to elicit a corpus of referring expressions from scenes that encouraged the use of spatial relations, and in which a range of alternatives for referential overspecification was available, including highly salient properties that were not strictly required for disambiguation.
The corpus contains 884 descriptions produced by 56 participants, and it was intended as a resource suitable to challenge preferences that are well-established in REG, such as the preference for colour (Pechmann, 1989), for absoluteness (van Gompel et al., 2014) and for shorter (e.g., atomic) descriptions over longer (e.g., relational) ones (Kelleher and Costello, 2009; dos Santos Silva and Paraboni, 2015). Given the relative domain complexity and proneness to referential overspecification, we assume Stars2 to be a likely useful resource for our present study on time-constrained reference production.

[2]For other possible triggers to referential overspecification, see for instance (Paraboni et al., 2017b).

3. Time-constrained human reference production

In this section we describe a simple psycholinguistic experiment in which human participants were engaged in a time-constrained reference production task, and we compare the elicited (time-constrained) descriptions with 'normal' descriptions available from the Stars2 corpus (cf. previous section).
The experiment was implemented as a game-like application in which participants were instructed to imagine that they were in charge of a highly critical, time-constrained task. To this end, we made use of a modified version of the Stars2 experiment setting in which participants had a limited time to complete the task, and were constantly reminded of the need to be quick.
The general hypothesis to be investigated is that referring expressions produced in time-constrained communication tend to present certain differences in length and attribute choice when compared to those produced under normal conditions.
Regarding description length, we assume that the effects of time-constrained communication may be comparable to those observed in critical situations discussed in (Arts et al., 2011). In other words, we expect that referring expressions produced under time-constrained conditions to be on average longer (and possibly overspecified) if compared to those produced under normal conditions.
Regarding attribute choice, we notice that studies as in (Pechmann, 1989) argue that *colour* is frequently selected even when not necessary for disambiguation, and for that reason we assume that the use of colour may be more frequent in time-constrained communication as well. Moreover, studies as in (van Gompel et al., 2014) suggest that relative *size* attributes (e.g., 'large', 'greater than', etc.) require considerable cognitive effort from speakers and hearers alike, and for that reason we assume that the use of *size* may be less frequent in time-constrained communication.
To verify these claims, we compared the semantic contents of a set of situations of reference with and without time constraint. This analysis takes into account six variables associated with a given description:

- The amount of information *length* represented by the number of annotated properties.
- The number of spatial relations *rel.count*.
- The degree of overspecification of the target object *over.tg* represented by the number of properties that could be removed from the description with no risk of ambiguity.
- The degree of overspecification of the first landmark object *over.lm1*.

- The overall frequency of the target colour attribute *colour.use*.
- The overall frequency of the target size attribute *size.use*.

3.1. Procedure

Participants of an in-person experiment similar to (Paraboni et al., 2017a) were briefed on the task, and were requested to produce descriptions for each target object to be identified in a sequence of images provided as quickly as possible in order to avoid a (fictitious) bomb explosion (and hence lose the game). Throughout the experiment, time constraint was reinforced with messages as in 'Hurry up, you need to describe all scenes or your time will run out!'.

3.2. Participants

23 volunteers with similar background as the Stars2 participants, with normal or corrected vision, being 12 (52%) female, and on average 32-years old.

3.3. Materials

13 image sequences from the Stars2 data collection task[3], in the same order observed in the construction of the corpus. Each sequence consisted of 20 images similar to Figure 2, being 4 fillers and 16 research stimuli. As in (Paraboni et al., 2017a), half of the stimuli represented contexts that favour the use of spatial relations as in the present example, and the other half did not, hence favouring the use of atomic description as in 'the red ball'.

3.4. Results

A series of 23 individual trials was performed upon appointment with each participant. As a result, a set of 368 descriptions of interest was obtained. Descriptions were subsequently annotated with their semantic properties by two annotators according to the annotation scheme adopted in (Paraboni et al., 2017a).

For the six variables of interest (*length*, *rel.count*, *over.tg*, *over.lm1*, *colour.use*, and *size.use*), Table 1 reports mean, variance and number of items n under consideration in both normal (i.e., as occurring in the Stars2 corpus) and time-constrained (the present experiment) conditions. The value of n corresponds to the total number of collected descriptions (368) except for the tests that distinguish relational and atomic conditions (*over.tg* and *over.lm1*), in which case n corresponds to half the data (184 descriptions each).

3.5. Discussion

We performed individual between-subjects ANOVA to compare the variables of interest in each condition with and without time restriction. Our findings revealed two significant differences discussed below. Other differences were not significant, an outcome that be explained by the small size of our data set.

First, we notice that time-constrained descriptions are, on average, more overspecified (*over.tg*) than those produced in normal situations. The difference is significant ($F(1, 101) = 15.947, p < 0.05$). This result is arguably consistent with the study in (Arts et al., 2011) regarding the production of referring expressions under critical conditions.

Second, we notice that time-constrained descriptions convey, on average, more references to colour (*colour.use*) than those produced in normal situations. The difference is significant ($F(1, 101) = 6.039, p < 0.05$). This result may be explained by the observation that, in the Stars2 domain, the use of colour information tends to be redundant, and possibly correlated with the overspecification of the target object itself. Thus, our findings for (*over.tg*) and (*colour.use*) may actually reflect the same phenomenon.

The presently annotated collection of time-constrained descriptions - hereby called Stars2T for analogy with the original, 'normal' dataset from (Paraboni et al., 2017a) - will be reused in the evaluation of a novel REG method described in the next section.

4. Time-constrained computational REG

As in the case of reference production in critical situations (Arts et al., 2011), the experiment described in the previous section suggests that time-constrained referring expressions may be more overspecified than those produced in so-called 'normal' circumstances. Based on this observation, this section describes a REG experiment involving a modified version of the Dale & Reiter algorithm (Dale and Reiter, 1995) in which the amount of referential overspecification is explicitly manipulated so as to generate descriptions that closely resemble those produced by human speakers in time-constrained situations.

4.1. Algorithms

The experiment makes use of a modified version of the Incremental approach (IA) (Dale and Reiter, 1995), hereby called IAL, that takes as an input an additional overspecification *Level* parameter to enable the generation of descriptions with a customisable amount of information. The value of *Level* is intended to represent the average size of descriptions in the relevant domain, and may in practice be computed from a small set of training examples as discussed in, e.g., (Koolen et al., 2012).

As in the original Incremental approach, IAL selects properties to compose an output description L by iterating over a domain-dependent list of preferred attributes P. At each step, properties that have some discriminatory power are selected until a uniquely identifying description is obtained, or until all properties in P have been attempted.

The difference between the Incremental approach and IAL is the stop condition. Once a suitable description has been obtained, the Incremental approach terminates. IAL, by contrast, uses the value of *Level* as a stop condition. If the number of selected properties in the output description L is equal to or greater than the expected *Level*, IAL also terminates. If not, then additional discriminatory properties will be added to L by following the list of preferred attributes P until the expected *Level* is reached.

To further illustrate the role of referential overspecification in time-constrained REG, we will also consider an implementation of the *Greedy* algorithm in (Dale, 2002) as an

[3]Randomly selected from 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.

	length		rel.count		over.tg		over.lm1		colour.use		size.use	
	normal	tc	normal	tc	normal	tc	normal	tc	normal	tc	normal	tc
mean	3.14	3.34	0.90	0.94	0.52	0.79	0.74	0.73	0.18	0.26	0.19	0.20
var.	2.84	2.47	0.66	0.56	0.39	0.39	0.31	0.33	0.14	0.23	0.15	0.20
n	368	368	368	368	184	184	184	184	368	368	368	368

Table 1: Normal and time-constrained (tc) human reference production

additional baseline system. Briefly, this strategy always selects the property with greatest possible discriminatory power, leading to the generation of brief descriptions with little or no overspecification.

4.2. Hypothesis

We hypothesise that adding a certain amount of overspecified information to a description in order to reach the average description length for a given domain will lead to output descriptions that more closely resemble time-constrained human descriptions than those generated by the Incremental approach. This hypothesis will be tested by comparing IA and IAL descriptions with a reference set of time-constrained descriptions produced by human speakers (cf. previous section) while measuring Dice scores (Dice, 1945). We expect IAL to obtain, on average, higher Dice scores than IA.

4.3. Data

We use the Stars2T corpus of time-constrained descriptions described in the previous section. The corpus was systematically divided in training (276 descriptions) and test (92 descriptions) sets so that instances of descriptions produced by every participant, and referring to every possible context, were approximately balanced within each set.

4.4. Procedure

For both IA and IAL, the P input parameter was computed from training data based on attribute frequencies (in which the most frequent attribute is to be attempted first). In the case of IAL, the additional *Level* parameter was also computed from the same training data as the average description length for each context.

Both algorithms - and also the *Greedy* baseline - took as an input the same test data. Evaluation was carried out by comparing every generated description with its human counterpart available from the test corpus while computing Dice coefficients. In addition to that, overall accuracy and MASI scores (Passonneau, 2006) were computed for illustration purposes.

4.5. Results

Table 2 summarised the results for the three algorithms under consideration applied to the test data.

Algorithm	Acc.	Dice	MASI
Greedy	0.18	0.62	0.36
Incremental	0.22	0.66	0.40
IAL	0.33	0.72	0.48

Table 2: Time-constrained computational REG

4.6. Discussion

Results suggest that the proposed IAL method generally outperforms both IA and *Greedy*. A Wilcoxon test shows that the difference between IAL and the second best approach - the Incremental algorithm - is significant both in terms of Dice (W =- 1203, Z =-4.32, $p < 0.0001$) and MASI (W = - 1158, Z = -4.16, p <0.0001) coefficients. In time-constrained reference production, adding overspecified information up to a certain *Level* (as in IAL) outperforms standard REG (as in IA). This offers support to our research hypothesis.

5. Final remarks

This article described an experiment involving human subjects engaged in time-constrained reference production task, and proposed a novel algorithm for the generation of descriptions under these circumstances. The experiment suggested that, in time-constrained communication, human speakers produce expressions that are on average longer or more overspecified than those produced in 'normal' conditions. These differences were subsequently exploited by an algorithm that allows description length to be manipulated explicitly.

Our findings are in principle consistent with studies that address the related issue of reference production in critical situations of communication, and pave the way for the design of referential overspecification strategies that, unlike machine learning methods (Ferreira and Paraboni, 2014a), do not require large sets of linguistic examples on every domain under consideration as training data.

Since referential overspecification may in principle occur in any situation of communication (and not only when there is a time constraint), as future work we intend to refine and validate the current approach by using descriptions available from existing REG corpora, that is, by considering 'normal' situations of reference as well. In doing so, we expect to obtain a more general solution for the generation of human-like overspecified referring expressions.

The Stars2T corpus of semantically-annotated time-constrained referring expressions is freely available for research purposes upon request.

6. Acknowledgements

This work has been supported by grant # 2016/14223-0, São Paulo Research Foundation (FAPESP). The authors are also thankful to Alex G. J. Lan for his support regarding the data collection task.

7. Bibliographical References

Arts, A., Maes, A., Noordman, L., and Jansen, C. (2011). Overspecification in written instructions. *Linguistics*, 49(3):555–574.

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. (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.

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.

Gatt, A., Krahmer, E., van Gompel, R., and van Deemter, K. (2013). Production of referring expressions: Preference trumps discrimination. In *35th Meeting of the Cognitive Science Society*, pages 483–488.

Kelleher, J. D. and Costello, F. J. (2009). Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306.

Koolen, R., Krahmer, E., and Theune, M. (2012). Learning preferences for referring expression generation: Effects of domain, language and algorithm. In *Proceedings of INLG-2012*, pages 3–11.

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.

Paraboni, I. and van Deemter, K. (1999). Issues for the generation of document deixis. In *Procs. of workshop on Deixis, Demonstration and Deictic Belief in Multimedia Contexts, in association with the 11th European Summers School in Logic, Language and Information (esslli99)*, pages 44–48.

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):451–459.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *International Conference on Language Resources and Evaluation (LREC)*.

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

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

van Gompel, R., Gatt, A., Krahmer, E., and Deemter, K. V. (2014). Testing computational models of reference generation as models of human language production: The case of size contrast. In *RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*, Edinburgh, Scotland.