From the Virtual to the Real World: Referring to Objects in Real-World Spatial Scenes

Dimitra Gkatzia, Verena Rieser
Department of Computer Science
Heriot-Watt University
Edinburgh EH14 4AS , UK
{d.gkatzia,v.t.rieser}@hw.ac.uk

Phil Bartie
School of Natural Sciences
University of Stirling
Stirling FK9 4LA, UK
phil.bartie@stir.ac.uk

William Mackaness
School of GeoSciences
University of Edinburgh
Edinburgh EH8 9XP, UK
william.mackaness@ed.ac.uk

Abstract

Predicting the success of referring expressions (RE) is vital for real-world applications such as navigation systems. Traditionally, research has focused on studying Referring Expression Generation (REG) in virtual, controlled environments. In this paper, we describe a novel study of spatial references from real scenes rather than virtual. First, we investigate how humans describe objects in *open, uncontrolled* scenarios and compare our findings to those reported in virtual environments. We show that REs in real-world scenarios differ significantly to those in virtual worlds. Second, we propose a novel approach to quantifying image complexity when complete annotations are not present (e.g. due to poor object recognition capabitlities), and third, we present a model for success prediction of REs for objects in real scenes. Finally, we discuss implications for Natural Language Generation (NLG) systems and future directions.

1 Introduction

REG has attracted considerable interest in the NLG community over the past 20 years (Krahmer and van Deemter, 2011; Gatt et al., 2014). While initially, the standard evaluation metric for REG was human-likeness, as compared to human corpora similarity as in TUNA (Gatt et al., 2009), the field has moved on to evaluating REG effectiveness by measuring task success in virtual interactive environments (Byron et al., 2009; Gargett et al., 2010; Janarthanam et al., 2012). Virtual environments however eliminate real-world uncertainty, such object recognition errors or cluttered scenes. In this paper, we investigate whether the lessons learnt in virtual environments can be transferred to real-world scenes. We consider the case where we are uncertain about the scene itself, i.e. we assume that the complexity of the scene is hidden and we are interested in identifying a specific object, and thus our work differs from approaches that generate descriptions for images such as (Mitchell et al., 2012; Feng and Lapata, 2013; Yang et al., 2011; Yatskar et al., 2014).

Related work has focused on computer generated objects (van Deemter et al., 2006; Viethen and Dale, 2008), crafts (Mitchell et al., 2010), or small objects in a simple background (Mitchell et al., 2013a; FitzGerald et al., 2013). One notable exception is the recent work by Kazemzadeh et al. (2014), who investigate referring expressions of objects in "complex photographs of real-world cluttered scenes". They report that REs are heavily influenced by the object type. Here, we are interested in studying REs for visual objects in urban scenes. As the success of a RE is heavily dependent on the complexity of the scene as well as its linguistic features, we are interested in modelling and thus predicting the success of a RE.

Initially, this paper presents and analyses a novel, real-world corpus REAL (to be released) – "Referring Expression Anchored Language" (Section 2), and compares the findings to those reported in virtual worlds (Gargett et al., 2010). We then provide a detailed analysis of how syntactic and semantic features contribute to the success of REs (Sections 4.1, 4.2, 4.3), accounting for unobservable latent variables, such as the complexity of the visual scene (as described in Section 3). Finally, we summarise our work and discuss the implications of our work for NLG systems (Section 5). The dataset and models will be released.

2 The REAL Corpus

The REAL corpus contains a collection of images of real-world urban scenes (Fig. 1) together with verbal descriptions of target objects (see Fig. 2)

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1936–1942,
Lisbon, Portugal, 17-21 September 2015.

Figure 1: Original picture.

Figure 2: Target object in yellow box.

Figure 3: The identified object by the validators.

generated by humans, paired with data on how successful other people were able to identify the same object based on these descriptions (Fig. 3). The data was collected through a web-based interface. The images were taken in Edinburgh (Scotland, UK), very early one summer morning. This was necessary to reduce the occlusion of city objects from buses and crowds, and to minimise lighting and weather variations between images.

2.1 Experimental Setup

There were 190 participants recruited (age between 16 to 71). Each participant was presented with an urban image (Fig. 1), where the target object was outlined by a yellow box (Fig. 2), and was asked to describe the target using free text. After completing a (self-specified) number of tasks, participants were then asked to validate descriptions provided by other participants by clicking on the object using previously unseen images (Fig. 3).

# participants	190
# images/ stimuli	32
# descriptions	868
# verifications	2618
– ambiguous	201
– not found	75
– correct	1994
– incorrect	251
– NA	7

Table 1: The REAL corpus

Overall, 868 descriptions across 32 images were collected, averaging around 27 descriptions per image. The balance of generation and validations was adjusted to ensure that all descriptions were identified by at least 3 other participants, generating 2618 image tag verifications. All cases were manually checked to determine if the 'correct' (green) or 'incorrect' (red) target had been identified Fig. 3. Overall, 76.2% of human descriptions provided were successfully identified. For the experiments reported in following sections, we summarised answers categorised as 'incorrect', 'ambitious' and 'not found' as *unsuccessful*.

2.2 Comparison to GIVE-2 Corpus

We now compare this data with human data generated for the GIVE-2 challenge (Gargett et al., 2010). In GIVE-2, the target objects have distinct attributes, such as colour and position. For instance, an effective RE in GIVE-2 could be "*the third button from the second row*". In real-world situations though, object properties are less well defined, making a finite set of pre-defined qualities unfeasible. Consider, for instance, the building highlighted in Figure 2, for which the following descriptions were collected:

1. *The Austrian looking white house with the dark wooden beams at the water side.*
2. *The white building with the x-shape balconies. It seems it's new.*
3. *The white building with the balconies by the river.*
4. *Apartments with balconies.*
5. *The nearest house on right side. It's black and white.*
6. *The white and black building on the far right, it has lots of triangles in its design.*
7. *The rightmost house with white walls and wood finishings.*

It is evident that the REAL users refer to a variety of object qualities. We observe that all participants refer to the colour of the building (*white, black and white, greyish-whitish*) and some mention location (*by the river, at the water side*).

Experimental Factors influencing Task Performance: In REAL, task success is defined as the ability to correctly identify an object, whereas in GIVE-2, task success refers to the successful completion of the navigation task. In contrast to GIVE-2, not all REAL participants were able to correctly identify the referred objects (76.2% task

	GIVE-2		REAL
	German	English	
Overall task success	100%	100%	76.2%
Task success (female)	100%	100%	78.8%
Task success (male)	100%	100%	69.6%
Length of descriptions (no. words)	5.2	4.7	16.01
Length of descriptions (female)	NA	NA	97.36
Length of descriptions (male)	NA	NA	91.38

Table 2: Descriptive statistics for GIVE-2 and REAL

success). We assume that this is because GIVE-2 was an interactive setup, where the participants were able to engage in a clarification dialogue. **Gender:** In REAL, gender was not a significant factor with respect to task success (Mann-Whitney U test, $p = 0.2$). **Length of REs (no. words):** In REAL, females tend to provide lengthier REs than males, however the difference is not statistically significant (Mann-Whitney U test, $p = 0.58$). In GIVE-2, only German females produced significantly longer descriptions than their male counterparts. **Relation between length (no. words) and task success:** The REAL data shows a positive relationship between length and success rate, i.e. for a one word increase in length, the odds of *correct* object identification is significantly increased ($p < 0.05$, Logit), i.e. longer and more complex sentences lead to more successful REs.

3 Quantifying the Image Complexity

We assume that the complexity of the urban scene represented in the image is hidden due to the lack of semantic annotations. Our dataset does not include any quantifiable image descriptions, such as computer vision output as in (Mitchell et al., 2012) or manual annotations as in (Yatskar et al., 2014). In addition, the same RE might not always result in successful identification of an object due to scene complexity. In order to marginalise the effect of the scene complexity, we exploit the multiple available data points per image. This allows us to estimate the average success rate of each referring expression $\overline{SR}_{RE}$ (the proportion of successful validations) and the average success rate of each image $\overline{SR}_i$ (the proportion of the correctly identified objects in the image). We use $\overline{SR}_i$ to marginalise over the (hidden) image complexity, where we assume that some pictures are inherently more complex than others and thus achieve lower success rates. Similar normalisation methods are used for user ratings to account for the fact that some users are more "tolerant" and in general give higher ratings (Jin and Si, 2004). We employ *Gaussian normalisation* (Resnick et al., 1994) to normalise image success rates by considering the following factors:

1. *Shift of average success rate per image:* some images are inherently easier than others and gain higher success rates, independently of the REs used. This factor can be accounted by subtracting average success rates of all images from the average rating of a specific image x.
2. *Different ratings:* there are 27 REs per image on average, some of which are harder to understand than others, thus they gain lower success rates. To account for this, the success rates of each image are divided by the overall SR variance.

The normalised image success rate (NSR_i) per image x is defined by the following equation:

$$NSR_i(x) = \frac{SR_i(x) - \overline{SR_i}}{\sqrt{\sum_n (SR_i(x) - \overline{SR_i})^2}} \quad (1)$$

Using the (NSR_i), we now investigate the REs in terms of their linguistic properties, including automatically annotated syntactic features and manually annotated semantic features.

4 Modelling REG Success

Unlike previous work, we use both successful and unsuccessful REs in order to build a model that is able to predict the success or the failure of a RE.

4.1 Syntactic Analysis of REG Success

We use the Stanford CoreNLP tool (Manning et al., 2014) to syntactically annotate the REs and we investigate which linguistic features contribute to the RE success in relation to the image complexity. Note that these analyses are based on normalised values, as discussed in Section 3).

Predicting RE Success Rate ($\overline{SR}_{RE}$): Initially, we compare successful and unsuccessful REs by taking the upper and lower quartiles and extracting their syntactic features., i.e. the top and bottom 25% of REs with respect to their average success rate, and group them into two groups. We then extract syntactic features of these two groups and compare their frequencies (occurrence per RE), means, and standard deviations (Table 3), and compare them using a t-test ($p < 0.05$). The difference between successful and unsuccessful ex-

	Successful REs			Successful REs		
	Mean	SD	Freq.	Mean	SD	Freq.
NP*	7.35	3.958	100	6.7	3.8	100
VP	1.45	1.673	41.8	1.46	1.923	58
PRN	.02	.181	2.1	.03	.193	2.4
NNP*	.57	1.131	27	.38	.918	19.3
NN*	4.2	2.284	98.8	3.79	2.441	98.1
DT	2.59	1.791	86.7	2.63	1.813	85.4
JJ*	1.92	1.61	80.9	1.66	1.288	81.1
CC	.4	.645	32.8	.31	.588	25
PP	2.52	1.754	92.3	2.54	1.778	85.8
ADJP	.2	.536	16.2	.041	.597	19.3
ADVP	.27	.538	22.8	.25	.539	19.8
RB	.34	.639	26.6	.34	.859	21.2
VBN*	.22	.465	20.3	.31	.445	10.8
NNS	.61	.902	40.7	.72	.782	53.3
CD	.27	.552	22	.25	.478	23.6

Table 3: Statistics regarding the linguistic features in successful vs unsuccessful referring expressions. (* denotes significant difference at $p < 0.05$).

pressions lies in the use of NP (Noun phrases), NNP (Proper noun, singular), NN (Noun, singular or mass), JJ (Adjective) and VBN (Verb, past participle) (Table 3). Successful REs include more NPs, including NNPs and NNs, which indicates that more than one reference is used to describe and distinguish a target object. This could mean that distractors are explicitly mentioned and eliminated or that the object of interest has a complex appearance, as opposed to simply structured objects, such as buttons, in GIVE-2. For example, the following description refers to a complex object:

> *The **large American-style wooden building** with **balcony painted cream and red/brown**. **Ground floor** is a **cafe with tables and parasols** outside.*

In addition, successful REs contain significantly more adjectives and verbs in past participle[1], which indicates that the object was further described and distinguished using its attributes, as for instance the following description:

> *Large modern **glass fronted** building, **butted up** against traditional Victorian terrace, slightly **set back** from road and with facing **bowed** frontage.*

The main difference between successful and unsuccessful REs is the amount of detail provided to describe and distinguish the target object. This is also in-line with our previous results that success is positively correlated to the number of words used (Section 2.2) and it might explain why humans overspecify.

[1]A participle is a form of a verb that is used in a sentence as modifier, and thus plays a role similar to that of an adjective or adverb, such as *built* or *worn*.

Models	R^2
Syntactic: NP+PP+ADVP+CD+length	.15
Semantic model: taxonomic + absolute	.338
Joint model: PP + taxonomic + absolute	.407

Table 4: Models and their fit.

To further verify this hypothesis, we build a predictive model of average success rate, using *multiple step-wise linear regression* with syntactic features as predictors. We find a significant ($p < 0.05$) positive relationship between success rate and NP, PP (Prepositional phrase), ADVP (Adverbial phrase), CD (Cardinal number), and length (Table 4). NPs are used to distinguish and describe the target object. ADVPs and PPs serve a similar function to adjectives in this case, i.e. to describe further attributes, especially spatial ones, like "the one near the river", "next to the yellow building". Cardinal numbers are used to refer to complex structured features of the target object, e.g. *two-story building* or *two large double doors*.

Predicting Image Success Rate ($\overline{NSR}_{image}$): We repeat a similar analysis for estimating how syntactic features relate to image success rate, i.e. how the image complexity, as estimated from the success rate of an image, influences how humans describe the target object, i.e. how human generated descriptions change with respect to the image complexity as estimated from the (normalised) success rate of an image. We find that humans use significantly more PPs and number of words ($p < 0.05$) when describing complex images.

In sum, syntactical features, which further describe and distinguish the target object (such as NPs, ADJ, and ADVPs and PPs) indicate successful REs. However, they cannot fully answer the question of "what makes a RE successful", therefore we enrich our feature set using manually annotated semantic features.

4.2 Semantic Analysis of REG

We extract semantic features by annotating spatial frames of reference as described in (Gargett et al., 2010). We annotate a sample of the corpus (100 instances), which allows us to perform a direct comparison between the two corpora.

Comparison to GIVE-2 Corpus: We observe that in the REAL corpus, the *taxonomic* property, the *relative property* and the *macro-level landmark*

Spatial Frame	REAL	GIVE-2	
		German	English
Taxonomic Property	**92***	53.66	58.51
Absolute Property	**57***	85.37	92.53
Relative Property	**15***	6.83	4.56
Viewer-centred	15	15.61	12.45
Micro-level landmark intrinsic	**9***	13.17	17.84
Distractor Intrinsic	**5***	10.73	14.11
Macro-level landmark intrinsic	**43***	6.83	4.15
Deduction by elimination	1	0.98	3.32

Table 5: Frequency of semantic frames in REAL vs. GIVE-2 (* denotes significant differences at $p < 0.05$, χ^2 test).

intrinsic property of the object in question are used significantly more often than in the GIVE-2 corpus (Table 5)[2].

In contrast, in GIVE-2 the *absolute property* of the object, such as the colour, and references to *distractors* are used significantly more often than in REAL. These results reflect the fact that scenes in REAL were more complex, and as such, relative properties to other objects and landmarks were used more often. In GIVE-2, target objects were mostly buttons, therefore, absolute descriptions (*"the blue button"*) or referring to an intrinsic distractor (*"the red button next to the green"*) are more frequent. In addition, real-world environments are dynamic. Humans choose to refer to immovable objects (*macro-level landmarks*) more often than in closed-world environments. In GIVE-2, immovable objects are limited to walls, ceilings or floors, whereas in REAL there is a wide range of immovable objects /landmarks that a user can refer to, e.g. another building, rivers, parks, shops, etc. Landmark descriptions will play an important role in future navigation systems (Kandangath and Tu, 2015).

Predicting RE Success Rate ($\overline{SR}_{RE}$): Next, we analyse which spatial frames significantly contribute to task success, using *multiple step-wise linear regression*.We find that *taxonomic* and *absolute* properties significantly ($p < 0.05$) contribute to the success of a referring expression (Table 4). Semantic features explain more of the variance observed in $\overline{SR}_{RE}$, than syntactic features.

[2]Note that for GIVE-2 we consider both, the German and the English data.

4.3 Joint Model of REG Success

Both syntactic and semantic features contribute to the success of a RE. Therefore, we construct a joint model for predicting $\overline{SR}_{RE}$ using *step-wise linear regression* over the joint feature space. We find that both syntactic and semantic features significantly ($p < 0.05$) contribute to the success of a RE, see Table 4. This model explains almost half of the variation observed in $\overline{SR}_{RE}$ ($R^2 = .407$). Clarke et al. (2013) reports an influence of visual salience on REG, therefore, in future, we will investigate the influence of visual features.

5 Discussion and Conclusions

From the results presented, the following conclusions can be drawn for real-world NLG systems. Firstly, semantic features have a bigger impact on the success rate of REs than syntactic features, i.e. content selection is more important than surface realisation for REG. Secondly, semantic features such as *taxonomic* and *absolute* properties can significantly contribute to RE success. Taxonomic properties refer to the type of target object, and in general depend on the local knowledge of the information giver. Similarly, the success of the RE will depend on the expertise of the information follower. As such, modelling the user's level of knowledge (Janarthenam et al., 2011) and stylistic differences (Di Fabbrizio et al., 2008) is crucial. Absolute properties refer to object attributes, such as colour. Attribute selection for REG has attracted a considerable amount of attention, therefore it would be interesting to investigate how these automatic attribute selection algorithms perform in real-world, interactive environments. Finally, the more complex scenes seem to justify longer and more complex descriptions. As such, there is an underlying trade-off which needs to be optimised, e.g. following the generation framework described in (Rieser et al., 2014).

In future, we will compare existing REG algorithms on our dataset, in a similar experiment to Mitchell et al. (2013b). Then, we will extend existing algorithms to take into account other properties such as material (e.g. "wooden"), components of the referred object (e.g. "balconies") etc. Finally, we will incorparate such an algorithm in interactive settings to investigate the influence of user dialogue behaviour and the influence of visual features, such as salience (Clarke et al., 2013), in order to improve the fit of our predictive model.

Acknowledgments

This research received funding from the EPSRC GUI project Generation for Uncertain Information (EP/L026775/1). The data has been collected through the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216594 (SPACEBOOK project).

References

Donna Byron, Alexander Koller, Kristina Striegnitz, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2009. Report on the First NLG Challenge on Generating Instructions in Virtual Environment (GIVE). In *12th European Workshop in Natural Language Generation (ENLG)*.

Alasdair D.F. Clarke, Micha Elsner, and Hannah Rohde. 2013. Where's wally: The influence of visual salience on referring expression generation. *Frontiers in Psychology*, 4(329).

Giuseppe Di Fabbrizio, Amanda Stent, and Srinivas Bangalore. 2008. Referring expression generation using speaker-based attribute selection and trainable realization. In *5th International Natural Language Generation Conference (INLG)*.

Yansong Feng and Mirella Lapata. 2013. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812.

Nicholas FitzGerald, Yoan Artzi, and Luke Zettlemoyer. 2013. Learning distributions over logical forms for referring expression generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Kristina Striegnitz. 2010. The GIVE-2 Corpus of Giving Instructions in Virtual Environments. In *7th International Conference on Language Resources and Evaluation (LREC)*.

Albert Gatt, Anja Belz, and Eric Kow. 2009. The TUNA-REG Challenge 2009: Overview and Evaluation Results. In *12th European Workshop in Natural Language Generation (ENLG)*.

Albert Gatt, Emiel Krahmer, and Kees van Deemter. 2014. Models and empirical data for the production of referring expressions. *Language, Cognition and Neuroscience*, 29(8):899 – 911.

Srinivasan Janarthanam, Xingkun Liu, and Oliver Lemon. 2012. A web-based evaluation framework for spatial instruction-giving systems. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*.

Srinivasan Janarthenam, Helen Hastie, Oliver Lemon, and Xingkun Liu. 2011. "the day after the day after tomorrow?" a machine learning approach to adaptive temporal expression generation: training and evaluation with real users. In *12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Rong Jin and Luo Si. 2004. A study of methods for normalizing user ratings in collaborative filtering. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 568–569, New York, NY, USA. ACM.

Anil Kandangath and Xiaoyuan Tu. 2015. Humanized navigation instructions for mapping applications. US Patent, 04.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Emiel Krahmer and Kees van Deemter. 2011. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, 2015/02/20.

Christopher Manning, Mihai Surdeanu, John Finkel, Jenny Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In *6th International Natural Language Generation Conference (INLG)*.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé, III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 747–756.

Margaret Mitchell, Ehud Reiter, and Kees van Deemter. 2013a. Typicality and object reference. In *Cognitive Science (CogSci)*.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2013b. Generating expressions that refer to visible objects. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstorm, and John Riedl. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In *ACM Conference on Computer Supported Cooperative Work (CSCW)*.

Verena Rieser, Oliver Lemon, and Simon Keizer. 2014. Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(5):979–994, May.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *4th International Natutral Language Generation Conference*.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *5th International Natural Language Generation Conference (INLG)*.

Yezhou Yang, Ching Lik Teo, Hal Daumé, III, and Yannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mark Yatskar, Michel Galley, Lucy Vanderwende, and Luke Zettlemoyer. 2014. See no evil, say no evil: Description generation from densely labeled images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*.