Applying Machine Learning to the Choice of Size Modifiers

Margaret Mitchell (m.mitchell@abdn.ac.uk)
Kees van Deemter (k.vdeemter@abdn.ac.uk)
Ehud Reiter (e.reiter@abdn.ac.uk)
Computing Science Department, University of Aberdeen
Aberdeen, Scotland, U.K.

Abstract

People use different size modifiers to refer to different object sizes; "the long table" is likely to be a different table from "the small table". However, the details influencing the selection of size modifier have not yet been uncovered. When is something "long", and when is something "small"? We introduce a connection between the visible dimensions of objects and the kinds of language people use to refer to them. First, we conduct an experiment to elicit size-denoting modifiers from images of real world objects. We find that we are able to effectively model the relationship between dimensional features and modifier choice using decision trees. The images are then used as input to an object segmentation algorithm, and we compare how well we can predict speakers' behavior using the real world measurements of the pictured objects and the image pixel-based measurements. We find that real world measurements are the best predictors of modifier choice, suggesting that people infer real world size features from images. However, automatically extracted pixel measurements do perform relatively well at predicting modifier choice, offering a potential connection between computer vision and natural language. When speaker identity is taken into account, modifier choice can be predicted with even greater accuracy (around 75%), and the difference between automatically extracted and real world measurements is no longer significant.

Introduction

Advances in the fields of natural language generation, image processing and computer vision have begun to make it possible to model the connection between visual properties and language. Given a set of images and corresponding descriptions, we can create a mapping between image features and language features (Kulkarni et al., 2011). These language features can then be analyzed by a natural language generation (NLG) system to produce human-like expressions.

We introduce preliminary work on such a system, focusing on generating reference to an object's SIZE. Examining the SIZE property in isolation gives us a tractable problem to solve within the larger problem of moving from visual input to natural language output. This is also in line with current research in the generation of referring expressions (GRE), where many algorithms are built to explicitly handle a SIZE property (Dale & Reiter, 1995; Krahmer, van Erk, & Verleg, 2003). Psycholinguistic research has also demonstrated that size is a primary aspect of natural reference to objects (Landau & Jackendoff, 1993; Sedivy, 2003; Brown-Schmidt & Tanenhaus, 2006).

One complication to this approach is that current philosophy in GRE treats an input property as a single feature and does not provide mechanisms for reasoning about how a property may involve interacting features. In Dale and Reiter (1995) and Krahmer et al. (2003), the knowledge base must mark elements as `large` or `small`. van Deemter (2006) treats size as a gradable property, producing size adjectives from numerical measurements (e.g., size = 33cm). None of these proposals do justice to the fact that size can involve a *combination* of dimensions; a turtle may be fat, or big, but seldom tall. Addressing this issue allows us to connect the visible dimensions of objects to size language. We can analyze how features related to an object's height and width predict size modifier choice, and apply this to a GRE algorithm.

To find out how an algorithm should choose between different size-denoting modifiers based on the heights and widths of objects, we conducted an experiment to elicit descriptions of real world objects (Mitchell, van Deemter, & Reiter, 2011). The present paper builds a discriminative machine learning model from the resulting corpus. The input to the model is the height and width of each object, and the output is the type of size modifier to generate. The size types predicted include a width/height type, corresponding to surface forms such as "tall" and "thin", an overall size type, corresponding to surface forms such as "big" and "small", and a type for expressions without size modification (e.g., "the square brownie").

We compare inputs to the model based on real world measurements, image pixel measurements extracted by hand, and image pixel measurements extracted using the semi-supervised SIOX algorithm (Friedland, Jantz, & Rojas, 2005). The semi-supervised approach connects modifier choice to the output of an image processing/computer vision technique known as object segmentation, providing a possible link between natural language and computer vision. We find that this approach works well, with an accuracy of 64.95% on unseen test images, but does not perform as well as the models built from real world measurements, which reach 69.44% accuracy. By adding speaker label as a model feature, accuracy from all models improves above 75%, and the difference between the semi-supervised approach and the real world approach is no longer significant.

We use a decision tree classifier in order to visualize how different features affect the selection of size type. Features that emerge with high information gain may be useful in a hand-coded GRE algorithm, and we walk through these details in the Results section. The trees built with speaker label also provide a concrete model of speaker variation for this task, and we outline the different speaker clusters the model uncovers.

This paper therefore makes three primary contributions: (1) a connection between the visual features of a scene and the generation of natural size language; (2) an exploration of visual features that may be useful in further work on human-like GRE; and (3) a model of speaker-dependent variation for the SIZE attribute. Both the images and elicited expressions are available at http://www.csd.abdn.ac.uk/~mitchema/corpora/size.html.

Background and Motivation

Size modifiers are used to refer to dimensional properties of objects. A modifier like "big" tends to be used in cases where an object is large in either two or all three of its dimensions, while modifiers like "thick" and "thin" may be applied when an object extends in a single dimension (Landau & Jackendoff, 1993). Size modifiers are common in visual scenes (van Deemter, van der Sluis, & Gatt, 2006; Viethen & Dale, 2009), and are especially prevalent when an object needs to be distinguished from another object of the same type (Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Brown-Schmidt & Tanenhaus, 2006). There is some evidence that the selection of size modifier can be predicted by using a machine learning approach that reasons about dimensional features extracted from a scene, such as an object's surface area (Roy, 2002).

Previous work on determining the form of an object description using machine learning has created models that predict a wide range of properties, such as the inclusion of color, location, etc., as well as the overall form of the noun phrase (e.g., personal pronoun, definite description). These approaches utilize a variety of contextual features, such as intentional influences and conceptual pact features (Jordan & Walker, 2005) and syntactic, semantic, and discourse features (Poesio, 2000).

Such work does not address the fact that different speakers will generate different reference within the same context (Reiter & Sripada, 2002), which is a large factor when the goal is to generate natural reference. This variation speaks to the need for models that can incorporate individual speaker profiles in generation.

In light of this, recent work in GRE has begun to use speaker-specific constraints in order to improve the performance of reference algorithms (Fabbrizio, Stent, & Bangalore, 2008). In work most closely related to the current study, Viethen and Dale (2010) use a decision tree classifier to predict the set of attributes different speakers will use to refer to geometric shapes. The results are mixed, largely due to the lack of data for many of the proposed classes; however, there is a significant increase in accuracy when speaker identity is included as a model feature.

In the current approach, we examine how well a small set of objective visual features perform at predicting the type of size modifier selected to refer to everyday objects. We include the size-based features of surface area and height-to-width ratio suggested by Roy (2002) to be correlated with distinct size adjectives. In contrast to earlier work on machine learning for generating object descriptions, the images are of real objects, the features do not rely on detailed annotation,[1] and the set of predicted classes is kept small. This narrows the machine learning task from earlier related work and avoids data sparsity issues. At the same time, it provides a relatively clear connection between the size aspects of a scene, such as the height and width of a target object, and natural referring expression generation.

It is important to note that at both ends of this connection, the problem is reduced to basic levels. The visual input of images is an obvious application for computer vision that utilizes object recognition. However, object recognition can only return regions of an image where an object is likely to exist, not the specific details of the object's dimensions (Walther, Itti, Riesenhuber, Poggio, & Koch, 2002; Lowe, 2004). To reason about an object's shape, an object segmentation approach is needed, with the general location of the object already specified. Work linking object recognition to object segmentation is still quite new (e.g., Zheng, Yuille, and Tu (2010)). Our approach therefore compares real world measurements to measurements extracted from semi-supervised object segmentation.

At the other end of the vision-language connection is GRE, a well-developed subfield within natural language generation. However, GRE has focused on categorizing which subset of scene attributes may be selected to identify an object. In this paper, we take a more fine-grained approach by exploring the use of a single attribute – SIZE – and several of its possible forms. We hope that this research provides a basic foundation from which to raise the complexity at both ends.

Procedure

Experiment

We manipulated the height and width of boards, books, brownies, and sponges. Each object appeared to the right of a comparator object of the same type (see Figure 1), and could appear in 24 different sizes, systematically varied along height and width axes: larger (++, axis 5/4 size of comparator), a little larger (+, axis 11/10 size of comparator), no difference (0, axis same size as comparator), a little smaller (-, axis 10/11 size of comparator) and smaller (- -, axis 4/5 size of comparator). A total of 96 images were used for this study, split in a Latin square design among three groups. Further details on this experiment are provided in Mitchell et al. (2011). For this paper, we report on results for an additional set of 414 participants.

For each expression, we annotate the modifiers as picking out *individuating* axes (I) – words like "tall" and "thin" – *overall* axes (O) – words like "big" and "small" – or *none* (N). These serve as the class labels for each image-based feature vector in the training data. Example expressions are given in Figure 2. The full list of size-denoting words for each class label is given in Table 1. Inter-annotator agreement on a randomly selected 10% of this data is high, Cohen's $\kappa = 0.94$.[2]

[1] In the semi-supervised approach we discuss, the features are extracted from images, but the ability to recognize such features in a scene is limited by how well an object segmentation algorithm works; we control this aspect by looking at clear, uncluttered scenes.

[2] 729 size modifiers were compared for the agreement score; 5 modifiers only labeled by one annotator are excluded.

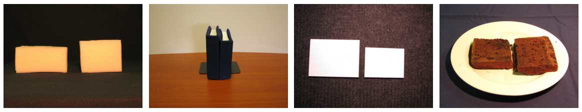

Figure 1: Example stimuli.

Object, Cond.	Expression	Class
sponges, h+w- -	taller sponge	I
boards, h- -w++	the shorter and slightly wider board with a diagonal top side	I
boards, h- -w- -	smaller board	O
brownies, h+ w-	the most square brownie	N

Figure 2: Example expressions for different <object, condition> stimuli. Conditions are composed of different measurements of the height (h) and width (w) axes.

Object Segmentation

In addition to the measurements of the real height and width of the objects, we measure the objects' height and width in image pixels. We also extract such information using the SIOX algorithm (Friedland et al., 2005), a semi-supervised method for object segmentation. We explain this algorithm briefly here.

The input for the SIOX algorithm consists of three user specified regions of a given image: known background, unknown region, and known foreground. To notate each region, we manually outline a general selection of the location of each object. The outer region of this selection becomes the known background, and the inner region the unknown region.

By selecting (brushing over) parts of the object, we specify the known foreground. Both known regions are then used in a classification task to identify which sections of the unknown region are background and which are foreground. The resulting output is an outline of the segmented object, separated from the surrounding background.

We then store each of the segmented objects as separate images. With this in place, an image processing tool can be used to extract pixel height and pixel width of each object image. We use CONJURE for this, a command-line based program implemented within ImageMagick (Cristy, Thyssen, & Weinhaus, 2011). Figure 3 shows an example of an image and extracted objects.

Machine Learning

Each of the 96 images represent an <object, condition> stimulus with associated features. There are a variety of size-based visual features available from the heights and widths extracted from the input images, listed in Table 2. These include Referent Features, features of the target object alone; Comparator Features, features of the comparator object on the left; and Comparison features, features that store the difference between the referent and comparator. These features may serve as the training/testing data in a machine learning approach where the class label in each instance corresponds to the size type (I, O, or N) used by a particular speaker for a particular image. The classification problem is therefore to use the visual features to predict the size type used by each speaker for each image.

Table 1: Root words for size modifier labels.

Label	Count	Vocabulary
I	3307	breadth, broad, deep, elongated, fat, flat, height, high, length, long, low, narrow, short, skinny, slender, slim, squat, stout, tall, thick, thin, wide, width
O	2614	big, large, little, shrunk, slight, small
N	703	- -

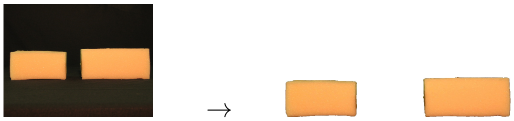

Figure 3: Example of original and extracted objects.

We use C4.5 decision tree classifiers as implemented within Weka (Hall et al., 2009) with default parameter settings. Performance is evaluated using leave-one-out validation, where the set of results from all speakers for each <object, condition> stimulus (each image) is tested against a model trained on all other objects and conditions.

Results

Results are presented in Table 3, listed as the percentage of correct predictions, and in italics, the percentage of testing folds where the predicted type was found in the majority of responses. We compare results based on the three kinds of visual measurements:

1. Automatically extracted image measurements (Auto): the pixel measurements extracted from the segmented objects within the pictures.

2. Gold-standard image measurements (Gold): pixel measurements measured by hand from the objects within the pictures.

3. Real World measurements (Real): the actual measurements of the pictured objects.

Accuracy is computed as the number of correct classifications divided by the number of classified instances, over all testing folds. If n is a testing fold in the set of testing folds N, t_i is the true class label of each instance i, and p_i is the predicted class label, then:

$$\text{Accuracy} = \frac{\sum_{i \in n \in N} (p_i = t_i)}{\sum_{n \in N} |n|}$$

Accuracy based on the automatically extracted pixel measurements indicates how well the system connecting object segmentation to reference generation performs. Accuracy based on gold standard and real world measurements provide

Table 2: Visual features extracted from images.

#	ID	Description
Referent Features		
1	type	object type
2	h	height of target
3	w	width of target
4	ratio	target height:width
5	surfar	surface area of target
6	hwdf	target height - target width
Comparator Features		
7	dh	height of comparator
8	dw	width of comparator
9	drat	comparator height:width
10	dhdwdf	comparator height - comparator width
Comparison Features		
11	hdhdf	target height - comparator height
12	wdwdf	target width - comparator width
13	ratdf	target ratio - comparator ratio

Table 3: Accuracy across folds.

	Auto	**Gold**	**Real**	**Oracle**	**Baseline**
Without	64.95%	62.80%	69.44%	75.88%	49.93%
Speaker	*65.63%*	*65.63%*	*75.00%*	*88.54%*	*47.92%*
With	75.33%	77.20 %	76.95%	100%	64.05%
Speaker	*91.67%*	*96.88%*	*95.83%*	*100%*	*71.88%*

a comparison indicating how well the system performs when the size data is provided manually.

The system connecting object segmentation to natural reference generation (Auto) performs relatively well, predicting 64.95% of response types. The comparison pixel measurement system (Gold) predicts 62.80% of response types, which is not significantly different from the automated approach (paired t-test, p=0.4104).

Interestingly, even though real world measurements may not be clear in photographs, we find that classification based on these measurements performs significantly better than classification based on the manually or automatically derived pixel measurements (paired t-test, real vs. auto: p=0.0363, real vs. gold: p=.0188). This suggests that people are good at reasoning about size in the real world from a two-dimensional image, and the connection between what a computer can see and what it can talk about may be improved with more sophisticated techniques for geometric reasoning.

Since all testing instances in each fold are identical, differing only in class label (the size type), we implement an oracle method to understand the upper bound of this task. This predicts the most common size type in each testing fold, which yields 75.88% accuracy. The results can be compared against a majority baseline that predicts the most common type from the training data in each fold. Without speaker, the majority class is always I, which is used in 3,307 of the 6,624 instances; 2,614 are O, and 703 are N.

Figure 4 A shows an Auto model built over all data, without speaker labels. We see that *dhdwdf*, the feature for the difference between the comparator's height and width, is selected as having the highest information gain. In other words, the learning approach finds that first splitting up the data based on the value of this feature is the optimal way to distinguish between different size choices; when the model sees a set of visual features for a size choice it has to guess, it will first check whether *dhdwdf* is less than or equal to -47 pixels.

In this model, the features related to *ratio* appear as strong predictors of size type. Both the height-to-width ratio of the referent object and the difference in height-to-width ratio between the referent and comparator object are used early on in the trees. This means that features of the target referent itself, as well as features derived from the comparison between referent and comparator, play a role in which label is selected. It also suggests that there may be a relationship between the selected size type and how close the height and width of the target object are to one another – for example, when the dimensions are far apart, individuating size modifiers may be preferred, resulting in expressions with words like "tall" and "thin", but when closer together (more square-shaped), overall size modifiers may be preferred, resulting in expressions with words like "big" and "small". Further testing is necessary to understand whether the behavior of these features is reflective of human use of these features.

It is interesting that the models are not composed entirely of comparison features, but incorporate features of the referent in isolation, such as its ratio and width. This runs counter to much work in GRE, where algorithms usually select features of a referent object based solely on comparison with features of surrounding objects. This data suggest there may also be a benefit in reasoning about the relationship between individual features of the referent object itself before surface realization.

Speaker-Specific Reference Generation

We next add speaker label as a feature in the data and evaluate how well the classifiers perform. This provides a way to distinguish between instances within each testing fold. The trees built using this feature also provide a model of speaker variation.

As shown in Table 3, accuracy improves, and this is significant for all three learned models (Auto, Gold, and Real, $p < .001$). These models outperform a majority baseline that predicts the majority size type used by each speaker based on the training data in each fold. The Auto models predict 75.33% of the observed size types, and predict the majority type for a testing fold 91.67% of the time. This is not significantly different from the predictions made by the Real models (paired t-test, $t = 1.685, p = 0.095$). The resulting trees have very low depth, tuning decisions to each speaker and then using a small set of individualized features to decide the final size type (Figure 4 B). Clear clusters emerge in this approach, producing a concrete model of speaker variation. Clusters with more than two speakers are given in Table 4.

A.

```
dhdwdf ≤ -47
|   ratdf ≤ -0.097
|   |   wdwdf ≤ 4
|   |   |   type = books : O
|   |   |   type = boards : O
|   |   |   type = brownies : O
|   |   |   type = sponges : I
|   |   wdwdf > 4
|   |   |   hdhdf ≤ 7: I
|   |   |   hdhdf > 7: O
|   ratdf > -0.097
|   |   ratdf ≤ 0.143
|   |   |   ratio ≤ 0.705
|   |   |   |   dhdwdf ≤ -49
|   |   |   |   |   ratio ≤ 0.688
|   |   |   |   |   |   surfar ≤ 41004
|   |   |   |   |   |   |   hdhdf ≤ -14: O
|   |   |   |   |   |   |   hdhdf > -14: I
|   |   |   |   |   |   surfar > 41004: O
|   |   |   |   |   ratio > 0.688: I
|   |   |   |   dhdwdf > -49: N
|   |   |   ratio > 0.705: O
|   |   ratdf > 0.143
|   |   |   ratdf ≤ 0.152: I
|   |   |   ratdf > 0.152
|   |   |   |   w ≤ 177: O
|   |   |   |   w > 177
|   |   |   |   |   w ≤ 238
|   |   |   |   |   |   hdhdf ≤ 6: O
|   |   |   |   |   |   hdhdf > 6
|   |   |   |   |   |   |   ratdf ≤ 0.245: N
|   |   |   |   |   |   |   ratdf > 0.245
|   |   |   |   |   |   |   |   hdhdf ≤ 22: N
|   |   |   |   |   |   |   |   hdhdf > 22: I
|   |   |   |   |   w > 238: O
dhdwdf > -47: I
```

B.

```
dhdwdf ≤ -47
|   ratio ≤ 0.674
|   |   spkr = A2E : 2d
|   |   spkr = A2J
|   |   |   hwdf ≤ -113: 1d
|   |   |   hwdf > -113: 2d
|   |   spkr = A2F : 1d
|   |   spkr = A32
|   |   |   ratio ≤ 0.561: 2d
|   |   |   ratio > 0.561: 1d
|   |   spkr = A2T : 1d
|   |   spkr = AW5 : 2d
|   |   spkr = A37 : 2d
|   |   spkr = A3G : 1d
|   |   spkr = A94
|   |   |   hwdf ≤ -113: 1d
|   |   |   hwdf > -113: 2d
|   |   spkr = A3U
|   |   |   dw ≤ 205: N
|   |   |   dw > 205
|   |   |   |   hdhdf ≤ 8: I
|   |   |   |   hdhdf > 8: O
|   |   spkr = AN3
|   |   |   wdwdf ≤ 35: 2d
|   |   |   wdwdf > 35: 1d
|   |   spkr = A34 : 2d
|   |   spkr = A1I : 1d
|   |   spkr = A35 : 2d
|   |   spkr = A2S
|   |   |   dhdwdf ≤ -126: 1d
|   |   |   dhdwdf > -126: 2d
|   |   spkr = A19
|   |   |   h ≤ 152: 1d
|   |   |   h > 152: 2d
|   |   spkr = A3I : 1d
|   |   spkr = A18 : 2d
```

Figure 4: Pixel-based decision tree without speaker labels (A) and a section of pixel-based tree with speaker labels (B).

Discussion

Generating human-like reference to visible, real world objects is possible by reconstructing the problem of GRE: Rather than analyzing the SIZE property as a single dimension in feature space (<SIZE:`large`>), it can be analyzed as a multi-dimensional property (<SIZE:[`height:y width:x ratio:z...`] >). In this way, output from a visual analysis may serve as input to a model that selects the most reasonable value (including *none*) for the given attribute.

Without speaker labels, the models built on real world measurements perform better than the models built on pixel image measurements. This suggests that a connection between language generation and object segmentation can be improved by adding a mechanism to reason about how the two-dimensional image space maps to a three-dimensional real world space.

The models built here point the way to further psycholinguistic work, such as research uncovering other factors that affect the modifier choice made by people (perhaps, for example, cognitive load). The features selected by the decision trees cluster speakers into several groups, and whether these features reflect the features humans use when referring to size is an area for future research.

Table 4: Speaker clusters, with count of speakers using each feature set to select size type.

Cluster	Feature Set				Count
1	dhdwdf	ratio	ratdf		265
2	dhdwdf	ratio	ratdf	w	36
3	dhdwdf	h	ratio	ratdf	23
4	dhdwdf	dw	ratio	ratdf	16
5	dhdwdf	dh	ratio	ratdf	16
6	dhdwdf	hwdf	ratio	ratdf	12
7	dhdwdf	hdhdf	ratio	ratdf	9
8	dhdwdf	ratio	ratdf	type	9
9	dhdwdf	ratio	ratdf	wdwdf	7

Conclusions

We have presented a way to generate natural reference based on visual input, focusing on the property of SIZE. This work illustrates that the generation of natural size modification may be possible utilizing two abstract size types, *individuating* size and *overall* size. By reasoning over a set of features instead of a single value for the SIZE attribute, models for reference generation can be built from basic information provided by a visual system.

We find that generating natural reference may be aided by reasoning about features of the referent in isolation, as well as by comparing the referent to other items in the scene. Features related to the height-to-width *ratio* of objects play a key role in predicting size type, and this may be useful for a hand-coded algorithm that aims to generate natural reference.

Taking speaker into account significantly improves accuracy, with the models building decisions for individual speakers. A model built from the entire dataset provides a classification of the speaker-dependent variation used in this domain, and we find that speakers can be neatly clustered into 9 main groups based on the dimensional features that best predict size modifier preference. This suggests that generating human-like language can be improved by building models for individual speakers. In a system that generates natural language, these models can be constructed as speaker 'profiles' that follow different language behavior, and such profiles can be built, for example, by clustering speakers in the training data together.

In future work, we aim to expand the kinds of size language the models predict, specifying more detailed classes within the two broad size types. We also plan to develop this approach to work with more sophisticated visual input. Computer vision techniques provide rich information on many visible features, such as COLOR, MATERIAL, ORIENTATION and TEXTURE, and using these features along with SIZE features will allow for the generation of more complex natural expressions. We hope that continuing research on the kinds of features that image processing and computer vision provide and the kinds of language that people produce will help to connect a computer's vision to its language.

Acknowledgements

This research was supported by the Scottish Informatics and Computer Science Alliance (SICSA) and the Overseas Research Students Awards Scheme (ORSAS). We thank Ellen Gurman Bard and Brian Roark for their help with this work.

References

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*, 592–609.

Cristy, J., Thyssen, A., & Weinhaus, F. (2011). *Imagemagick.* GNU General Public License. (http://www.imagemagick.org/www/conjure.html)

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

Fabbrizio, G. D., Stent, A. J., & Bangalore, S. (2008). Trainable speaker-based referring expression generation. *Proceedings of the 12th Conference on Computational Natural Language Learning*, 151–158.

Friedland, G., Jantz, K., & Rojas, R. (2005). SIOX: simple interactive object extraction in still images. *Proceedings of the Seventh IEEE International Symposium on Multimedia*, 253–259.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations*, *11*(1).

Jordan, P. W., & Walker, M. A. (2005). Learning content selection rules for generating descriptions in dialogue. *Journal of Artificial Intelligence Research*, *24*, 157–194.

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

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., et al. (2011). Baby talk: Understanding and generating image descriptions. *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, *16*, 217–265.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*, 91–110.

Mitchell, M., van Deemter, K., & Reiter, E. (2011). On the use of size modifiers when referring to visible objects. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Poesio, M. (2000). Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. *Proceedings of the 2nd Language Resources and Evaluation Conference, LREC-2000*, 211–218.

Reiter, E., & Sripada, S. (2002). Human variation and lexical choice. *Computational Linguistics*, *28*, 545–553.

Roy, D. K. (2002). Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, *16*, 353–385.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*, 3–23.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*, 109–147.

van Deemter, K. (2006). Generating referring expressions that involve gradable properties. *Computational Linguistics*, *32*(2), 195–222.

van Deemter, K., van der Sluis, I., & Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. *Proceedings of the 4th International Conference on Natural Language Generation*.

Viethen, J., & Dale, R. (2009). Referring expression generation: What can we learn from human data? *Proceedings of the PRE-CogSci Workshop*.

Viethen, J., & Dale, R. (2010). Speaker-dependent variation in content selection for referring expression generation. *Proceedings of the 8th Australasian Language Technology Workshop*, 81–89.

Walther, D., Itti, L., Riesenhuber, M., Poggio, T., & Koch, C. (2002). Attentional selection for object recognition - a gentle way. *Proceedings of the 2nd Workshop on Biologically Motivated Computer Vision*, 472–479.

Zheng, S., Yuille, A., & Tu, Z. (2010). Detecting object boundaries using low-, middle-, and high-level information. *Journal of Computer Vision and Image Understanding*, *114*(19), 1055–1067.