# Visual and linguistic predictors for the definiteness of referring expressions

**Manjuan Duan, Micha Elsner** and **Marie-Catherine de Marneffe**
Linguistics Department
The Ohio State University
Columbus, OH, 43210, USA
{duan,melsner,mcdm}@ling.osu.edu

## Abstract

This study explores the impact of visual context on the conceptual salience of a discourse entity, using descriptions of how to find specific targets in cartoon scenes. Significant positive correlation is observed between larger and more salient objects and definite expressions, whereas more cluttered images are positively related to indefinite expressions. Incorporating these findings with other linguistic factors, we build a mixed-effects logistic regression model for predicting referring forms. The model reaches 62% accuracy. This study helps us to understand better how physical context, like an image, determines the linguistic properties of a discourse.

## 1 Introduction

When presented with a picture, how do you start your description? How will visual factors affect the expressions you use? And how do these factors interact with contextual and discourse features? Answering these questions will help to build a connection between the visual clues we perceive from a picture and the particular linguistic expressions we choose to describe it. It will also facilitate referring expression generation (REG) (Krahmer and van Deemter, 2012): the task of generating natural and contextually proper referring expressions.

In this study, we examine the roles played by visual features of an object and its visual context in determining whether, in a description, it will be mentioned by a simple definite NP, a long, descriptive definite expression, an indefinite, a demonstrative or a pronoun. We find that visual features like area and low-level visual salience are positively associated with definite referring expressions as a whole, suggesting that visually prominent objects are treated as more conceptually salient when we describe them.

These results are important for two reasons. First, they draw a firm connection between linguistic theories of reference which appeal to salience, and the low-level perceptual mechanisms from which salience arises. By doing so, they help to situate these theories in the wider context of cognitive science. Secondly, there is comparatively little research investigating the effect of visual properties on referring forms. Most previous research on text generation for REG has focused on content selection (Krahmer and van Deemter, 2012) where several studies have found effects of visual salience. These results suggest that human speakers also take vision into account during sentence planning and realization.

We use mixed-effects regression models to analyze the importance of several visual and linguistic factors in a corpus of visual-scene referring expressions (Clarke et al., 2013). These models can be used to predict referring forms on new data. Our classifier achieves 62% accuracy, which is 30% better than the majority baseline, and 6% better than a classifier without visual features, demonstrating its usability for generating contextually appropriate referring expressions for visual scenes.

## 2 Background

### 2.1 Linguistic theory

Linguists have proposed many different theories to account for the relationship between the salience of discourse entities and the kinds of re-

*Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue, December 16-18, 2013, Amsterdam, The Netherlands.*

ferring expressions that can be used to describe them. Although different terms, such as topicality, givenness, accessibility, prominence, familiarity or salience, are used, they all converge on one point: referring expressions reflect the cognitive status of discourse entities they refer to (Prince, 1999; Chafe, 1976; Givon, 1983; Gundel et al., 1993; Ariel, 1988; Roberts, 2003).

Several of these theories match the cognitive or attentional states associated with a discourse entity and specific linguistic forms of reference. Gundel et al. (1993) use the term *givenness* to illustrate how salient a discourse entity is. They specify a scale of attentional states corresponding to different forms of referring expressions. The Givenness Hierarchy they suggest is: *in focus* > *activated* > *familiar* > *uniquely identifiable* > *referential* > *type identifiable*. Related to these six cognitive statuses are the forms of referring expressions that these statuses license: *it* > *that, this, this N* > *that N* > *the N*> *indefinite this N* > *a N*. Each status on the hierarchy is a necessary and sufficient condition for the appropriate use of a linguistic form. For example, a discourse entity has to be *in focus* to be referred by a pronoun, or a discourse entity has to be *uniquely identifiable* to license the definite expression *the N*.

Along the same line, Roberts (2003) proposes that the use of a definite NP presupposes that that the NP is *familiar* (i.e., that there is a corresponding discourse referent already in the discourse context), and that this discourse referent is *unique* among the discourse referents in the context. She also further differentiates *familiarity* into *strong familiarity* and *weak familiarity*. Strong familiarity is reserved for the more commonly assumed notion of familiarity, where it usually involves explicit previous mention of the entity in question, while an entity is weakly familiar when its existence is *entailed* by the local context. Hence, weak familiarity subsumes strong familiarity but is more inclusive, including discourse referents introduced non-linguistically, on the basis of contextual entailments (including perceptually accessed information) alone.

Ariel (1988; 1991) proposes a similar theory in which the complexity of referring forms reflects their accessible status in our mind. Basically, more reduced linguistic forms suggest more accessible or more salient status in the discourse. Based on her empirical study, she proposes a graded Accessibility Marking Scale in which she differentiates nominal descriptions with modifiers and those without modifiers. In general, expressions with modifiers refer to entities with lower accessibility. For example, short definite expressions denote discourse referents that are more accessible than long ones; the descriptive content of long definites helps to further single out their discourse referents.

These theories are attractive to us because they make efforts to capture the correlation between cognitive status on the one side and linguistic forms of referring expression on the other side.

What is generally missing is a fully grounded theory which explains how low-level percepts affect the cognitive status ranking. While it is universally acknowledged that non-linguistic factors play a role, most research has focused on linguistic features which can create or indicate a high cognitive status for an entity: for instance, Grosz et al. (1995) proposes a ranking scale of grammatical roles played by the discourse entities, *subject* > *object* > *others*, see also (Kameyama, 1986; Hudson et al., 1986). Other factors like the distance between the entity and its previous mention, the competition from other discourse entities and the (in)animacy of the discourse entities have also been studied as cues to determine the cognitive status of a discourse entity (Hobbs, 1976; Mitkov, 1998; Haghighi and Klein, 2010). When present, these linguistic features are highly influential, often overriding non-linguistic perceptual factors (Viethen et al., 2011a). But when they are not, less is known about which perceptual features matter in selecting appropriate referring forms.

### 2.2 Referring expression generation

Both psycholinguists and text generation specialists have examined precisely the case in which visual information has the greatest influence: one-shot referring tasks (i.e., without discourse context) involving an object in a visual scene.

Viethen et al. (2011b) analyze a corpus of map-task dialogues and find that visual context is *not* an important factor in deciding content of a referring expression, even for first mentions. However, other studies have found effects for visual features.

Kelleher et al. (2005) claim that salience—both visual and linguistic—is an important overarching semantic category structuring visually situated discourse. They describe a system which uses simple measurements of visual salience—bounding

box area and distance to screen center—for both language understanding and REG content selection, and find these features are helpful. Duckham et al. (2010) use a variety of visual and perceptual features to select landmarks for computer-generated navigation instructions.

Clarke et al. (2013) also find a role for visual features in content selection. (They argue that the discrepancy with Viethen et al. might be accounted for by the stimuli—the images Clarke et al. use are more complex.) They find that visual properties (salience, clutter, area, and distance) influence referring expression generation for targets embedded in images from "Where's Wally?" books. Referring expressions for large target objects are shorter than those for small targets, and expressions about targets in highly cluttered scenes use more words. Also, people are more likely to choose large, salient objects which are close to the target as landmarks in relational descriptions.

Comparatively fewer studies have investigated how low-level visual features affect linguistic forms. Montag and MacDonald (2011) examined how visual salience affects the linguistic structure choice in terms of passive or active voice in relative clauses.

Closer to our work, Vogels et al. (2013) study how visual salience affects the choice of referent and the choice of referring forms when interacting with linguistic context in two story-completion experiments. They find that visual salience influences the choice of referent and does so independently of linguistic salience. But visual salience does not affect the choice of referring forms, which are strongly affected by linguistic salience. They conclude that visual salience has an influence on the global interpretation of the scene, but does not directly affect the accessibility status of individual entities— that is, people use different types of information in choosing a referent and choosing a referring expression.

In contrast, we do find effects from visual information on referring form, but nonetheless, we believe our study accords with Vogels et al. (2013). In their study, the two possible linguistic forms considered are pronouns and full noun phrases. Pronouns are a referring form which is highly sensitive to linguistic context, and our results also show they are relatively insensitive to visual effects; our strongest effects are in distinguishing different types of NP. Moreover, our one-shot referring task provides no linguistic context to begin with, while the story completion task of Vogels et al. (2013) provides previous referring expressions for the entities in all experimental conditions.

All the research introduced above shows that salient landmarks are more likely to be chosen in route description or scene descriptions than less salient ones and salient objects are more likely to be chosen as subject referent, which establishes the important role that visual salience plays in content selection. Both Montag and MacDonald (2011) and Vogels et al. (2013) study how visual salience affect our choice of concrete linguistics forms, but these studies involve highly controlled experimental environments in which perceptual variables are manipulated in a fairly coarse way, so that visual salience can be considered as a categorical variable rather than a continuum. Moreover, although Vogels et al. (2013) considers the choice of pronouns vs NPs , they leave open the issue of definiteness: what kind of NP to produce.

In this paper, we reanalyze Clarke et al. (2013)'s data, investigating which visual features of an object in an image or visual properties of the image as a whole affect people's choice of concrete linguistic referring forms. This study not only reveals the effects of various perceptual factors but also quantifies their relative importance. We show that both visual characteristics of the referent (visual salience and size) and a characteristic of the image as a whole (clutter) correlate with increased use of definite expressions. Furthermore, since visual factors have measurable effects on people's choice of referring forms, then consideration of these factors in referring expression generation tasks should be beneficial.

### 2.3 Visual salience

The visual salience (*Salience*) of an object (Toet, 2011) is a description of how much the object stands out from the background. Perceptual psychologists have developed models of visual salience, which typically aggregate low-level features such as color and contrast, and compare the features around each point to those in the image in general in order to predict how different the point will look from its surroundings. The size and central location of an object are also important (Tatler, 2007). Such models can predict fixations during scene viewing (Itti and Koch, 2000). Re-

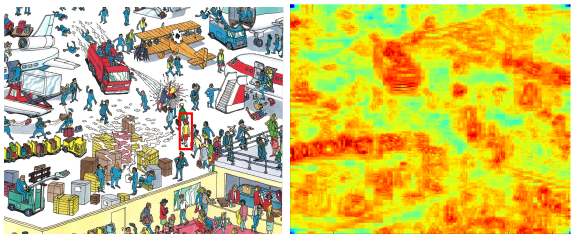


Figure 1: An image from our corpus and the corresponding visual salience map produced by the bottom-up component of Torralba et al. (2006); red indicates high salience scores, blue low salience scores.

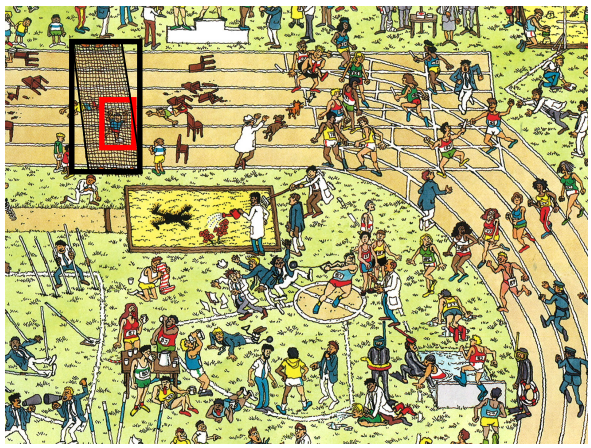
lated models from visual search (Wolfe, 1994) can also be used to predict how quickly subjects find a target object in a visual search task.

The Torralba et al. (2006) model used in our experiments is a typical contrast-based salience model (which we augment by including area, centrality and distance features as independent predictors).[1] It computes a visual salience score for each pixel in the image using a bank of oriented filters, then assigns a salience score to each bounding box which is the maximum over pixels it contains. The pixel scores are illustrated in Figure 1, which illustrates the visual prominence of the fire truck and the line of baggage handlers.

Visual clutter is a measurement of scene complexity; high clutter leads to difficulty when visually searching for objects (Henderson et al., 2009). Models of clutter (Rosenholtz et al., 2007) also depend on local image features such as color and orientation; in general, if these features are highly variable (many different colors and edge angles are represented), the scene will appear cluttered and hard to search.

## 3 Methods

We use a corpus collected in Clarke et al. (2013),[2] consisting of descriptions of specific target people in cartoon scenes from the children's book series "Where's Wally". The descriptions were elicited on Mechanical Turk, by asking participants to explain to someone else how to find a target person in the picture. Clarke et al. (2013) annotated the textual descriptions by marking references to visible objects and linked each one to a corresponding bounding box in the image. Their annotation scheme distinguishes two types of objects: the *target* is the person in the picture whom the subject was instructed to describe, while *landmarks* are other objects in the picture that the subject uses to describe the target. They also distinguish between textual mentions of landmarks that are part of a relative description ("near the bus") (Dale and Haddock, 1991), and those whose existence is *established* without giving a relative description ("look at the bus"). An example of the annotation is given in Figure 2.



```
Under <lmark rel="targ" obj="imgID">a
net</lmark> is <targ>a small
child wearing a blue shirt and red
shorts</targ>.
```

Figure 2: An example image and RE from the corpus with the target marked by a red box. The annotator has added a black box for the landmark (in this case the net). Words describing the target and landmark in the RE are XML-tagged.

Our goal here is to characterize how visual features affect the way people perceive definiteness of a discourse entity and choose referring forms accordingly from a cognitive/linguistic standpoint. We therefore used the totality of the descriptions in the corpus, without conducting experiments to determine whether they would lead to a successful/quick identification of the target by the listener. The fact that we did not filter out such "bad/unsuccessful" descriptions might be a weakness as far as applications are concerned, but from the cognitive/linguistic investigation that concerns us, these descriptions are a valuable source of information about how speakers compose descriptions.

[1] The Torralba et al. (2006) model also includes a top-down component which models task-based attentional effects, but this is not used.

[2] http://datashare.is.ed.ac.uk/handle/10283/336

| | Pron | Demo | SDef | LDef | Indef |
|---|---|---|---|---|---|
| Counts | 575 | 213 | 1013 | 1584 | 1594 |
| % | 11.5 | 4.3 | 20.3 | 31.8 | 32.0 |

Table 1: Distribution of referring forms.

We distinguish six classes of referring form: pronouns, demonstratives, short definite NPs, long definite NPs, indefinite NPs and bare singulars. We manually annotate each tagged mention of a visual object with its appropriate class.[3] Demonstratives are NPs headed by *this, that, these* and *those*. Definite NPs are those headed by *the*. Short definite NPs are definite NPs without any modifiers and long definite NPs are those with modifiers like adjectives, prepositional phrases, and relative clauses. We split the definites in this way in order to investigate the Accessibility Marking theory of Ariel (1988). Indefinite NPs are those headed by *a, an, some* or plural nouns. Bare singulars are singular nouns not headed by any determiners, like "man with a hat" or "brown dog"; these are ungrammatical in standard English, but occur in Mechanical Turk elicitations. The corpus contains 447 bare singulars; a preliminary analysis using the features below showed that these were similar in their distribution to definites and usually misclassified as such. We conclude that the bare singular form is an alternate form of the definite, and in the rest of our analysis one-word bare singulars are merged with short definite NPs and longer bare singulars with long definite NPs (Table 1).

We perform one-vs-all mixed-effects logistic regression analyses with R (Bates et al., 2011). We incorporate random intercepts for speaker (N=115) and image (N=11), and three types of fixed-effects features: task-based, visual and linguistic.

### Task-based features

The task features indicate whether the object being referred to is the target of the description (*Target*) or a landmark (*Lmark*).

### Visual features

Visual features of the described object include its area (*Area*) as well as its centroid-to-centroid distance from the target (*Distance*). Another feature captures whether its bounding box overlaps with that of the target or, if it is a landmark in a relative description of some other object, with that object (*Overlap*) (Kelleher et al., 2005; Golland et al., 2010).

We also use two models from the perception literature as features in our analysis. Both of them are previously-implemented models from the perceptual psychology literature. We use the values computed and distributed by Clarke et al. (2013), which measure the visual salience of bounding boxes by using the bottom-up component of Torralba et al. (2006). We also compute visual clutter using two models proposed in Rosenholtz et al. (2007).[4] Feature congestion (*Congestion*) measures the variance in features like different colors, orientations, or luminance contrast changes in a given local area. Sub-band entropy (*Clutter* or *Clt*) measure represents the intuition that an "organized" scene is less cluttered. With more organization, and thus more redundancy, the brain (or computer) can represent an image with more efficient encoding, thus a lower value in this measure. It is inversely related to how many bits could be saved by JPEG-compressing the image (Rosenholtz et al., 2007; Asher et al., 2013). All the values of visual features used in this paper are distributed as part of the corpus.

### Linguistic features

We use linguistic features found to be useful in previous studies of definiteness and information status (Nissim, 2006). In some cases we modified these feature definitions to rely on surface ordering rather than syntactic annotations, due to our lack of a parser for the Mechanical-Turk-elicited text.

*Coref*: We check if the phrase refers to a previously-mentioned entity, treating two phrases as coreferent if they resolve to the same bounding box in the image.

*Establish*: This feature captures whether the annotator marked the expression as *establishing* existence rather than part of a relative description, such as "look at the X", rather than a relative description like "near the X".

*There-be*: We have an explicit feature to capture *there+be* existential construction, known to disfa-

[3] The corpus also contains tags for non-visual objects ("the bottom left") and tags that are not mentions ("first on the left [implied *of X*]"); we exclude these from our analysis.

[4] We compute these scores ourselves, using the Matlab tools distributed by Rosenholtz.

| Features | Pron | Demo | SDef | LDef | (Def) | Indef |
|---|---|---|---|---|---|---|
| **Task** | | | | | | |
| Target | 1.44 ** | 3.46 *** | 0.60 * | -0.58 ** | -0.003 | -1.16 *** |
| Lmark | -0.74 · | 1.78 *** | 1.07 *** | -0.86 *** | -0.09 | 0.22 |
| **Linguistic** | | | | | | |
| Coref | 4.49 *** | 0.75 *** | 0.04 | -1.61 *** | -0.09 | -2.35 *** |
| There-be | -15.25 | -15.43 | -3.75 *** | -3.84 *** | -4.61 *** | 5.33 *** |
| Be | -3.33 *** | -3.01 ** | -2.11 *** | -2.77 *** | -2.99 *** | 3.88 *** |
| First | 0.89 *** | 0.14 | -0.50 ** | -0.31 ** | -0.54 *** | -0.41 ** |
| Prep | -0.13 | 0.01 | 0.01 | 0.16 ** | 0.28 *** | -0.38 *** |
| Establish | 0.55 * | 2.16 *** | -0.17 | -0.49 ** | -0.73 *** | 0.50 ** |
| **Visual** | | | | | | |
| Area | -0.35 · | -0.81 * | 0.64 *** | -0.38 *** | 0.63 *** | -0.67 *** |
| Salience | -0.26 ** | -0.01 | 0.08 | 0.05 | 0.11 * | -0.02 |
| Overlap | 0.001 | 0.47 · | 0.07 | -0.4 *** | -0.46 *** | 0.61 *** |
| Distance | 0.16 | -0.11 | 0.15 ** | 0.19 * | 0.37 *** | -0.66 *** |
| Clutter | 0.54 | -0.17 | 0.01 | -0.43 * | -0.37 * | 0.34 * |
| Congestion | 0.02 | -0.21 · | 0.001 | 0.07 | 0.07 | 0.01 |
| **Interaction** | | | | | | |
| Target:Clt | -0.59 | 0.19 | 0.04 | 0.36 ** | 0.42 ** | -0.37 * |
| Area:Clt | 0.09 | -0.54 * | -0.01 | -0.09 * | 0.39 *** | -0.47 *** |
| Salience:Clt | 0.05 | -0.05 | 0.28 *** | -0.07· | -0.11 * | 0.15 ** |

Table 2: Coefficients learned by the one-vs-all mixed-effects models for predicting referring forms. Significance codes: p-value $< 0.001$, $***$; p-value $< 0.01$, $**$; p-value $< 0.05$, $*$; p-value $< 0.1$, $\cdot$. The model includes all pairwise interactions, but only significant interactions are shown. The "Def" column shows coefficients for a merged class containing both long and short definites.

vor definites (Ward and Birner, 1995).

*Syntactic position*: We checked whether the target is directly preceded by any form of *to be* (*Be*); whether it is directly preceded by a preposition (*Prep*) or whether it appears sentence-initially, a proxy for the subject grammatical role (*First*).

## 4 Results and analysis

The coefficients from our one-vs-all mixed effects logistic regression analysis are shown in Table 2.[5] The linguistic features generally behave as the existing literature leads us to expect. A previous coreferent mention has the expected impact on referring forms (Roberts, 2003): pronouns and demonstratives are favored as indicated by the positive estimate for *Coref*, whereas indefinites are disfavored (negative coefficient). Indefinite NPs are positively associated with *There* and *Be*. Definite NPs are positively related to *Prep*, indicating that uniquely identifiable discourse entities are more likely to be the complements of prepositions. *First* is positively related to pronouns, which supports the hypothesis that back-looking centers like pronouns tend to appear at linguistically salient positions like subject position to achieve better discourse coherence (Grosz and Sidner, 1986).

As for visual features, we find main effects of *Area* in favor of short definite NPs, against long definites and strongly against indefinites. This result accords with the Accessibility Marking Scale proposed by Ariel (1988), which uses short definites for more accessible objects, then long definites and finally indefinites.

The results for *Salience* are smaller, but appear to be similar. Visual salience has non-significant positive associations with both short and long definites; if both classes of definite are analyzed together, the effect reaches significance. We suspect the failure to find it with either subgroup is due to reduced power because of the relatively smaller datasets. Overall the results confirm our hypothesis that larger and more visually salient objects are also perceived as more prominent and tend to be referred to by definite expressions, especially short definites.

*Overlap* is positively related to indefinite expressions and *Distance* is positively related to definite expressions. A closer look will show that these two measures are inversely related; usually, when two objects are overlapped, the centroid distance between them is short. In other words,

[5] We also considered the distance of an object to the center of the image, but its effect was not significant.

| Features | Accuracy | Sig vs. |
|---|---|---|
| Baseline (majority) | 32.01 | |
| Task features | 38.92 *** | baseline |
| Linguistic features | 54.68 *** | baseline |
| Visual features | 42.19 *** | baseline |
| Task + visual features | 43.30 *** | task |
| Task + ling features | 56.11 . | ling |
| Ling + visual features | 58.08 *** | ling |
| Task + ling + visual | 62.06 *** | ling + visual |

Table 3: Prediction results for the different feature types, with distinction between short and long definite referring expressions. The last column indicates whether results significantly differ (Mann-Whitney U test).

| Features | Accuracy | Sig vs. |
|---|---|---|
| Baseline (majority) | 51.7 | |
| Task features | 55.32 *** | baseline |
| Linguistic features | 72.44 *** | baseline |
| Visual features | 55.94 *** | baseline |
| Task + visual features | 56.78 *** | task |
| Task + ling features | 73.27 | ling |
| Ling + visual features | 74.15 | ling |
| Task + ling + visual | 74.74 | ling |

Table 4: Prediction results for the different feature types with short and long definite expressions combined. The last column indicates whether results significantly differ (Mann-Whitney U test).

speakers use more definite expressions to refer to objects far from the target of the description, while using more indefinites to refer to objects close by. Landmarks that are close by can be helpful even if they are hard to see (by helping the listener confirm that they have found the target). But distant landmarks must be easy to find in their own right, and this makes them better candidates for definite mentions.

Converging with the findings discussed above, the estimates for *Clutter* suggest that indefinite expressions are more likely to be used in a more crowded image. *Area* also interacts with *Clutter*: large objects are more likely to be definite and less likely to be indefinite when the image is more cluttered overall. This supports the results from linguistic research that indefinites need to be *type identifiable* (Gundel et al., 1993) while definites need to have uniquely identifiable referents (Gundel et al., 1993; Roberts, 2003). In an image where a lot of similar objects crowd together, many objects, especially smaller ones, will be hard to uniquely identify, so speakers may avoid using definite references for them. Alternatively, speakers might not be able to easily verify that the object is in fact unique in the image.

Using the predictions obtained from the five one-vs-all logistic regressions, we classify 479 randomly chosen NPs held out as test data, using the standard highest score strategy. Table 3 shows the classification accuracies. We find that all three types of features are significantly more effective than a majority baseline (always "indefinites"). Linguistic features are very robust in predicting referring forms as widely recognized by prior research, which itself improve the overall accuracy from 32% to 55%. Adding visual features also leads to significant improvement of predicting results on the top of baseline, linguistic features and task-based features, which gives stronger support for our hypothesis that low-level visual features play an important role in predicting linguistic forms for referring expressions. Our strongest model, using all feature sets together, scores 62%.

| Gold ↓ | Proposed → | | | | |
|---|---|---|---|---|---|
| | Pron | Demo | SDef | LDef | Indef |
| Pron | 454 | 17 | 32 | 62 | 10 |
| Demo | 49 | 26 | 28 | 108 | 2 |
| SDef | 44 | 11 | 398 | 463 | 97 |
| LDef | 63 | 3 | 157 | 1180 | 181 |
| Indef | 31 | 0 | 31 | 488 | 1044 |

Table 5: Confusion matrix for predicted referring forms.

Table 4 shows the classification accuracies when short and long definite expressions are combined. All three types of features are still significantly more effective than the baseline majority, now definites. However, adding visual features does not lead to a significant improvement on top of linguistic and task-based features. This means combining short and long definite expressions reduces the prediction of visual features, which suggest visual features are most effective in differentiating short and long definite expressions.

Table 5 shows per-category prediction results for each of the referring forms, cross-validated over the entire dataset. Most pronouns are predicted to be pronouns, despite their low percentage in our data (11%); 16% are labeled as definites, and less than 2% as indefinites. Very few demonstratives (12%) are correctly predicted,

since they are extremely under-represented in our data (4%). However, most of them are predicted as definites (64%) and pronouns (23%). 11% of definites are labeled as indefinite, showing that pronouns, demonstratives and definite expressions, as a group, share some common features, and our model draws a relatively sharp distinction between this group and the indefinites.

Although different cognitive states are proposed in linguistic research as necessary conditions for definite expressions, such as *uniquely identifiable* by Gundel et al. (1993) and *weak familiar* by Roberts (2003), all these theories claim that discourse entities which have higher cognitive status in the givenness scale, like *in focus* or *activated* can be referred to by either definite NPs or proforms like pronouns or demonstratives. We observe this predicted overlap in the usage of these three referring forms in the confusion results.

Of the remaining errors, we believe many are due to individual differences between speakers in terms of visual perception or describing style. Inspection of the random intercepts reveals that speakers vary in the overall proportions of different referring forms they use. In some cases this seems to be a matter of style: some people phrase their referring expressions as instructions ("Look for the man standing aside the red truck"), others describe ("A man standing...") and some use a telegraphic style ("man, in blue jeans, standing...").

Figure 3 also suggests that visual properties like "area" have different effects on people's choice of whether to use definite or indefinite expressions. Most subjects (lines curving sharply to the upper left) follow the general trend of using definite expressions for larger objects, but a few show weaker trends, or no trend at all. Whether the variance is caused by speakers perceiving the image differently, or reacting differently to visual factors, deserves future study.

## 5 Conclusion

In this study, we have revealed the correlation between the visual features of discourse entities and their referring forms. We find visual features like *area* and *salience* are positively related to definite expressions and indefinite expressions are more likely to be used in crowded images. Based on these findings, we train a classifier to predict the referring forms for these visual objects. Our classifier achieves 62% overall accuracy, 30% higher than the majority baseline. This study helps us to better grasp the interaction between linguistic properties of the discourse and the physical context in which utterances are grounded. In future work, we hope to incorporate these features into a full-scale surface realization system.

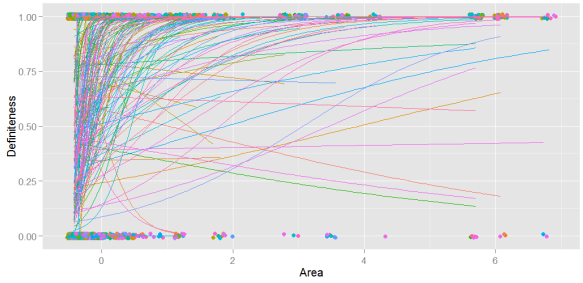



Figure 3: Logistic regression lines for proportion of definiteness as predicted by area for each of the 151 speakers in our data (data items shown as colored points). In general, larger area leads to more definite descriptions, but the effect varies across speakers and describing tasks.

## Acknowledgements

We thank Jefferson Barlew, Alasdair Clarke, Gregory Kierstead, Hannah Rohde, Craige Roberts and the reviewers for their useful comments on earlier drafts of this paper.

## References

Mira Ariel. 1988. Referring and accessibility. *Journal of Linguistics*, 24:65–87.

Mira Ariel. 1991. The function of accessibility in a theory of grammar. *Journal of Pragmatics*, 16:443–464.

Matthew F. Asher, David J. Tolhurst, Tom Troscianko, and Iain D. Gilchrist. 2013. Regional effects of clutter on human target detection performance. *Journal of Vision*, 13(5):1–15.

Doug Bates, Martin Maechler, and Ben Bolker. 2011. lme4: Linear mixed-effects models using S4 classes. Comprehensive R Archive Network.

Wallace Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics and point of view. In C. Li, editor, *Subject and Topic*. Academic Press, New York.

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 (Perception Science)*, Issue

on Scene Understanding: Behavioral and computational perspectives.

Robert Dale and Nicholas J. Haddock. 1991. Generating referring expressions involving relations. In *EACL*, pages 161–166.

Matt Duckham, Stephan Winter, and Michelle Robinson. 2010. Including landmarks in routing instructions. *Journal of Location Based Services*, 4(1):28–52.

Talmy Givon. 1983. Topic continuity in discourse: an introduction. In T. Givon, editor, *Topic continuity in discourse: a quantitative cross-language study*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Cambridge, MA, October. Association for Computational Linguistics.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307, June.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, Los Angeles, California, June. Association for Computational Linguistics.

John M. Henderson, Myriam Chanceaux, and Tim J. Smith. 2009. The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of vision*, 9(1):1–8.

Jerry R. Hobbs. 1976. Pronoun resolution. Technical Report 76-1, City College New York.

Susan B. Hudson, Michael K. Tanenhaus, and Gary S. Dell. 1986. The effects of the discourse center on the local coherence of a discourse. In *Program of the Eighth Annual Conference of the Cognitive Science Society*.

Laurent Itti and Christof Koch. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506.

Megumi Kameyama. 1986. A property-sharing constraint in centering. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*.

J. Kelleher, F. Costello, and J. van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167:62–102. Connecting Language to the World.

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

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 869–875, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Jessica Montag and Maryellen MacDonald. 2011. How visual salience affects structure choice: Implications for audience design. In *Poster presented at the 24th Annual CUNY Conference on Human Sentence Processing*, Stanford, CA.

Malvina Nissim. 2006. Learning information status of discourse entities. In *Proceedings of EMNLP*, pages 94–102, Morristown, NJ, USA. Association for Computational Linguistics.

Ellen Prince. 1999. How not to mark topics: 'topicalization' in English and Yiddish. In *Texas Linguistics Forum*. University of Texas, Austin.

Craige Roberts. 2003. Uniqueness in definite noun phrases. *Language and Philosophy*, 26:287–350.

Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. 2007. Measuring visual clutter. *Journal of Vision*, 7:1–21.

Ben W. Tatler. 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):1–17.

Alexander Toet. 2011. Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2131 – 2146.

Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. 2006. Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, 113:766–786.

Jette Viethen, Robert Dale, and Markus Guhe. 2011a. Generating subsequent reference in shared visual scenes: Computation vs re-use. In *Proceedings of*

*the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1158–1167, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Jette Viethen, Robert Dale, and Markus Guhe. 2011b. The impact of visual context on the content of referring expressions. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 44–52, Nancy, France. Association for Computational Linguistics.

Jorrig Vogels, Emiel Krahmer, and Alfons Maes. 2013. Who is where referred to how, and why? the influence of visual saliency on referent accessibility in spoken language production. *Language and Cognitive Processes*, 28(9):1323–1349.

Gregory Ward and Betty Birner. 1995. Definiteness and the English existential. *Language*, 71(4):722–742, December.

Jeremy M. Wolfe. 1994. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1:202–238.