Information Structure Prediction for Visual-world Referring Expressions

Micha Elsner
Department of Linguistics,
The Ohio State University
melsner@ling.osu.edu

Hannah Rohde
Linguistics & English Language,
University of Edinburgh
hannah.rohde@ed.ac.uk

Alasdair D. F. Clarke
School of Informatics,
University of Edinburgh
a.clarke@ed.ac.uk

Abstract

We investigate the order of mention for objects in relational descriptions in visual scenes. Existing work in the visual domain focuses on content selection for text generation and relies primarily on templates to generate surface realizations from underlying content choices. In contrast, we seek to clarify the influence of visual perception on the linguistic form (as opposed to the content) of descriptions, modeling the variation in and constraints on the surface orderings in a description. We find previously-unknown effects of the visual characteristics of objects; specifically, when a relational description involves a visually salient object, that object is more likely to be mentioned first. We conduct a detailed analysis of these patterns using logistic regression, and also train and evaluate a classifier. Our methods yield significant improvement in classification accuracy over a naive baseline.

1 Introduction

Visual-world referring expression generation (REG) is the task of instructing a listener how to find an object (the *target*) in a visual scene. In complicated scenes, people often produce relational descriptions, in which the target object is described relative to another (a *landmark*) (Viethen and Dale, 2008). While existing REG systems can generate relational descriptions, they tend to focus on content selection (that is, choosing an appropriate set of landmarks for each object). Surface realization (turning the selected content into a string of words) is handled by simple heuristics, such as sets of templates. Complex descriptions, however, have a non-trivial information structure— objects are not mentioned in an arbitrary order. Numerous studies in non-visual domains show that English speakers favor constructions that place familiar (given) information before unfamiliar (new) (Bresnan et al., 2007; Ward and Birner, 2001; Prince, 1981). We show that this pattern also holds for visual-world referring expressions (REs), and moreover, that objects with sufficient visual prominence are treated as given. Thus, we argue that the concept of salience used in surface realization should incorporate metrics from visual perception.

In this study, we create a model of information ordering in complex relational descriptions. Using a discriminative classifier, we learn to predict the information structuring strategies used in our corpus. We compare these strategies to the typical given/new pattern of English discourse. Experiments on a corpus of descriptions of cartoon people in the childrens' book "Where's Wally" (Handford, 1987), corpus described in (Clarke et al., 2013), show that our approach significantly outperforms a naive baseline, improving especially on prediction of non-canonical orderings.

This study has three main contributions. First, it demonstrates that humans use sophisticated information ordering strategies for REG, and therefore that the template strategies used in previous work do not adequately model human production. Second, it makes a practical proposal for an improved model which is capable of predicting these orderings; while this model is not a full-scale surface realizer, we view it as an important intermediate step towards one. Finally, it makes a theoretical contribution: By linking the information structures observed in the data to the existing re-

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 520–529,
Gothenburg, Sweden, April 26-30 2014.

search on salience and information structure, we show that visually prominent objects are treated as part of common ground despite the lack of previous mention.

2 Related work

Computational models of REG (Krahmer and van Deemter, 2012) focus mainly on content selection: Given a list of objects in the scene and their visual attributes, such models decide what information to include in a description so as to specify the target object. Early systems (with the exception of Dale and Haddock (1991)) did not produce relational descriptions. Nor did these systems model the visual salience of the objects or attributes under discussion.

Later models (Kelleher et al., 2005; Kelleher and Kruijff, 2006; Duckham et al., 2010) introduce simple models of visual salience, prompted by psycholinguistic research which shows that objects are more likely to be selected as landmarks when they are easy for an observer to find (Beun and Cremers, 1998). Clarke et al. (2013) extend these results with a more complicated model of visual salience (Torralba et al., 2006). Fang et al. (2013) similarly note that generated REs should avoid information that is perceptually expensive to obtain. However, these results focus on content selection rather than surface realization.

In comparison to selection, surface realization for REG has received little attention. Many researchers do not even perform realization, but simply compare their systems' selected content with the gold standard under metrics like the Dice coefficient. The TUNA challenges (Gatt et al., 2008; Gatt et al., 2009; Gatt and Belz, 2010) are an exception; participants were required to provide surface realizations, which were evaluated via NIST, BLEU and string edit distance. Many participants used a template-based realizer written by Irene Langkilde-Geary, which imposes a fixed ordering on attributes like "size" and "color" but has no provisions for relational descriptions. A few participants created their own realizers. Brugman et al. (2009) describe a system with multiple hand-written templates. Di Fabbrizio et al. (2008) propose several learning-based systems; the most effective were a dependency-based approach which learned precedence relationships between pairs of words, and a template-based approach which learned global orderings over sets of attributes. Neither approach is designed to handle relational descriptions, nor do they incorporate visual information. Duan et al. (2013), also studying the Wally corpus, demonstrates that visual features affect determiner choice for NPs, but do not study information structure.

Several studies give basic principles for information structure in English discourse. Prince (1981) introduces the key distinctions between discourse-old and new entities (previously mentioned vs not mentioned) and hearer-old and new entities (familiar to the listener vs not familiar). Clark and Wilkes-Gibbs (1986) extends the latter distinction to a notion of common ground; entities in the common ground are familiar to both participants in the discourse, and each participant is in turn aware of the other's familiarity. As noted by Prince (1981) and expanded on by Ward and Birner (2001) and in Centering Theory (Grosz et al., 1995), the first element in an English sentence is generally reserved for old information, while new information is usually placed at the end. For instance, see these (contrived) examples:

(1) a. **Obama** adopted **a dog named Bo**.
b. **#A dog named Bo** was adopted by **Obama**.

Ex. (1-a) demonstrates the standard order (under the assumption that *Obama* is familiar to a reader of this paper while *Bo* may not be). (1-b) violates the ordering principles and is likely to be judged less felicitous. Importantly, *Obama* is hearer-old not because of a preceding discourse mention but due to (assumed) general knowledge; it is an *unused* (Prince, 1981), or *existential* (Bean and Riloff, 1999) entity. General knowledge shared by speakers of a community is one way in which an entity enters the common ground. Along with this shared socio-cultural background, speakers may also share physical co-presence and linguistic co-presence (Clark, 1996). They can indicate salient entities, individuals, or entire events by engaging their listener in joint attention via pointing or gaze cueing (Baldwin, 1995; Carpenter et al., 1998); in this paper, we demonstrate that visual prominence is also sufficient.

Maienborn (2001) explicitly suggests that this topic-comment structure principle is the motivation for the frequent appearance of locative modifiers in clause-initial position; however, she gives no felicity conditions on *when* this leftward movement is expected. Since most of the modifiers in

this study are locatives, our data should be taken as endorsing this theoretical position, but supplying felicity conditions in terms of common ground.

These principles have been applied to computational surface realization in non-visual domains (Webber, 2004; Nakatsu and White, 2010, and others). Freer-word-order languages such as German also have predictable information structures which have been employed in surface realization systems, but these require a different structural analysis than in English (Zarrieß et al., 2012; Filippova and Strube, 2007).

3 Information structures in our corpus

In this section, we define the particular ordering strategies which we investigate in the rest of the paper. We begin by defining some terms: A relational description includes two objects, the *anchor*, which is the object being located, and the *landmark*, an object which is mentioned to make it easier to locate the anchor. The anchor may be the *target* of the entire expression, or it may in turn serve as a landmark in another relational description (as in "*the man* next to *the horse* next to *the building*" where "horse" serves as both a landmark for "man" and an anchor for "building".[1] The REs in this corpus reflect the variation in the way speakers constructed their descriptions: Some produced multiple complete sentences; others used abbreviated language and compacted their expression into a single sentence or phrase. In this paper we use the term "ordering" to refer to speakers' decisions of whether to precede or postpose a reference to one object relative to their reference to another. In this way, the "syntax" of the description is built out of references to particular objects (the noun phrases) and the relationships between those references. Note that the references may consist of a short phrase ("the man with the sword") or an entire clause ("he is standing and holding a sword")

In our corpus, speakers use three primary strategies to order anchors and landmarks, exemplified by the following REs from our corpus (shown with **bold** for text describing the anchor and *italics* for text for landmarks):

(2) Near the *hut that is burning*, there is a **man holding a lit torch in one hand, and a sword in the other**.

(3) **Man** closest to *the rear tyre of the van*.

(4) There is **a person standing** in *the water* **wearing a blue shirt and yellow hat**

Ex. (2) places the landmark so that it precedes the anchor; Ex. (3) shows the landmark following it. Ex. (4) shows a more complex structure, which we refer to as *interleaved*, where information about the anchor is given in multiple phrases and the landmark phrase appears between them.[2] (These orders are determined with respect to the first mention of the landmark.) We denote these ordering strategies as PRECEDE, FOLLOW and INTER respectively.

We also distinguish between landmarks which are only mentioned in relation to an anchor and those which are first introduced in a non-relative construction such as "look at the X" or "there's an X":

(5) There is *a horse rearing up on its hind legs*. Behind *the horse* is **a man laying down on his back completely flat and straight.**

Since these constructions establish the existence of a landmark without immediately incorporating it into the description, we denote these as ESTABLISH constructions.

Finally, our annotation scheme distinguishes between genuine landmarks (visible objects or groups of objects in the scene) and image *regions* like "the left" or "bottom center":

(6) *Bottom center*, **man looking left**

4 Dataset

We use a collection of referring expressions elicited on Mechanical Turk, previously described in (Clarke et al., 2013).[3] The dataset contains descriptions of targets in 11 images from the childrens' book *Where's Wally*[4] (Handford, 1987; Handford, 1988); in each image, 16 people were designated as targets. Each participant saw each scene only once. An example scene is shown in Figure 1. The participant was instructed to type a description of the person in the red box so that another person viewing the same scene (but without the box) would be able to find them; to make sure

[1] In our examples below, the anchor is the target of the overall expression, i.e., the intended referent in the REG task.

[2] This structure is not *syntactically* discontinuous, but visually it is; if the listener wants to confirm these details visually, they must first look at the person, then look away at the water and then look back at the person.

[3] Via `http://datashare.is.ed.ac.uk/handle/10283/336`

[4] Published in the USA as *Where's Waldo*.

this was clear, as part of the study instructions, they completed a few visual searches based on text descriptions. The image in the figure also contains a black box (not part of the initial stimulus), which the annotator has added to designate the landmark object "burning hut"). The dataset contains 1672 descriptions, contributed by 152 different participants (152 participants × 11 scenes).

The REs are annotated for visual and linguistic content. The annotation scheme indicates which substrings of the RE describe the target object, another mentioned object or an image region. References to parts or attributes of objects are not treated as separate objects; "a man holding torch and sword" in Figure 1 is a single object. The mentioned objects are linked to bounding boxes (or for very large objects, bounding polygons) in the image.

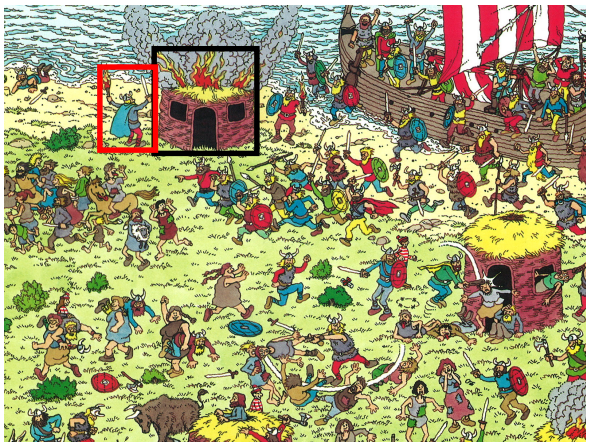
For each mention of a non-target object, the annotation indicates whether it is part of a relational description of a specific *anchor*, and if so which; if it is not, it receives an ESTABLISH tag. These annotations are used to determine the ordering strategies used in this study. In some cases, the linkage between objects is implicit:

(7) ...there are *4 men smoking*... the man you are looking for is **the one** [=of the 4 men] **leaning against a crate**

In the above RE, *4 men* is first introduced in an ESTABLISH construction. The word "one" refers implicitly to part of this set of men, so the annotator marks a relational link from "4 men" to "one". In our analysis in this study, we treat the entity "crates" as anchored to the target ("one") on the basis of this implicit link (so that this is an instance of the PRECEDE-ESTABLISH pattern), but we do not treat the hidden link itself as a mention or try to predict its nonexistent "position" in the string.

5 Distribution of ordering strategies

We first describe the distribution of these strategies across the corpus as a whole. As shown in Table 1, landmarks are ordered about equally to the FOLLOW or PRECEDE of the objects they help to locate. Regions, on the other hand, prefer the PRECEDE ordering. The INTER ordering is less common, but still quite well-represented. The ESTABLISH construction (initial "there is" or "look at") occurs only with PRECEDE ordering, and indeed can be viewed as a syntactic strategy for achieving such an order. We will explain these characteristic patterns in linguistic terms in Section 7.



`The <targ>man</targ> just to the left of the <lmark rel="targ" obj="imgID"> burning hut</lmark> <targ>holding a torch and a sword</targ>.`

Figure 1: Example scene (red box indicates target) with annotated referring expression. Words in `<targ>` tags describe the target. A single landmark (the burning hut, indicated by the `rel` attribute) is mentioned in a relational description whose anchor is the target; the annotator has marked it with a black box.

As in most discourse tasks (Ford and Olson, 1975; Pechmann, 2009), speakers display a fair amount of variability. To measure this, we examine each anchor/landmark pair which is mentioned by more than one speaker, and compute how often these speakers use the same strategy. There are 664 such pairs,[5] appearing a total of 2361 times in the corpus.[6] Of these, 66% agree on the directional strategy.[7] Separately, 14% of the expressions use an ESTABLISH construction, and 43% of these are agreed on by the majority. (The remaining variation could in principle have two sources: The content of the expression as a whole could affect the realization of a particular pair of objects, or individual speakers might simply differ in their usage patterns.) Nonetheless, there is a good deal of regularity in speakers' decisions. In the rest of the paper, we attempt to model and predict this regularity.

[5] 286 of these pairs are mentioned by exactly two speakers.

[6] This is more than the total number of referring expressions in the corpus, because many of the REs contain multiple pairs of entities.

[7] If strategies were assigned randomly using the overall marginals, we would expect only 34% agreement. Using this method of calculating chance agreement, we would obtain a Cohen's κ of .48.

	PRECEDE	INTER	FOLLOW
Region	60 (440)	21 (160)	19 (138)
L-mark	38 (977)	25 (632)	37 (945)
		ESTABLISH	NON-EST.
PRECEDE landmark		51 (495)	49 (482)

Table 1: Distribution of ordering strategies for all landmarks and regions in the corpus: *% (count)*. An additional 24 landmarks occur with no associated anchor (and therefore no discernible order).

6 Visual and non-visual information

Since visual properties are known to affect landmark selection (Kelleher et al., 2005; Viethen and Dale, 2008), we expect them to influence information structure as well. Our system uses three visual properties to predict information structure; we select properties that are known from previous work to help predict whether a landmark will be mentioned. These properties are the **area of the anchor and landmark**, the **distance** between them (Golland et al., 2010, among others) and their **centrality (centr.)** (distance from the center of the screen) (Kelleher et al., 2005).[8] These properties are all indicators of visual salience (Toet, 2011), the property which makes objects in a scene easy to find quickly (Wolfe, 2012) and tends to draw initial gaze fixations (Itti and Koch, 2000). We also include indicators for whether the anchor is the **target** object, and whether the landmark is an **image region (reg)** (see section 3).

In addition, we give a few non-visual features derived from the content structure. These include the **number of dependents** (landmarks which relate to each object in the description) and the **number of descendants** (the direct dependents, their dependents and so forth). When the speaker has to arrange a large number of landmarks, they tend to vary the ordering more, because of heavy-shift effects (White and Rajkumar, 2012) and the difficulty of preposing more than one constituent.

7 Regression analysis

To gain some insight into the influence of different features, we conduct a logistic regression analysis. For each pair of $(anchor, landmark)$ occurring in a relational description, we attempt to predict the manner of realization (direction and ESTABLISH). We performed a logistic regression for each class (one-vs-all); thus there are four regressors in total, making 0-1 predictions for PRECEDE, PRECEDE-ESTABLISH, INTER and FOLLOW.

Because their distributions are heavily skewed, area is transformed to square root area and distance/centrality values are log-transformed as in Clarke et al. (2013).[9] Features are scaled to zero mean and unit variance. Finally, centrality values are negated so that higher values indicate more central objects; this is for ease of interpretation. We fit models using random intercepts for speaker and image using the LME4 package (Bates et al., 2011), then removed all fixed effects which were never significant for any class and reran the analysis until a minimal model was reached (Crawley, 2007). This minimization removed the number of descendants features (but kept number of direct dependents). Table 2 shows the significant coefficients, standard deviations and Z-scores. (Note that as the regressions are separate, the coefficients are comparable reading down columns, but not across rows).

The regression analysis shows that as landmarks get larger, they are more likely to be realized with the PRECEDE ($\beta = 3.27$) or INTER ($\beta = 1.28$) strategies (but not PRECEDE-ESTABLISH) and less likely ($\beta = -3.76$) to be placed following. (This does not appear to be the case for landmarks that are central; these are slightly more likely to be ordered FOLLOW ($\beta = .81$).) The PRECEDE-ESTABLISH construction is neither favored nor disfavored by landmark area. It does, however, have a strong preference for landmarks with many dependents ($\beta = 2.38$), since these are more naturally realized in the clause-final position introduced by a "There is X"-type construction. In contrast, landmarks with many dependents disfavor the INTER strategy ($\beta = -1.07$), since this would require placing a heavy NP in a central rather than rightward position.

There are also a few effects of visual features of the anchor objects. Larger anchors (which are easier to see in their own right) prefer landmarks to FOLLOW ($\beta = .35$). This presumably reflects the fact that, since the listener is more likely to see them quickly, such anchors are more often re-

[8] Following Clarke et al. (2013), we attempted to also measuring distinctiveness from the background using a perceptual model of visual salience (Torralba et al., 2006). Although this measure is effective in predicting landmark selection, it proves uninformative here for predicting information structure, yielding no significant effects in any analyses.

[9] We use these continuous values in our analysis; our classifier model (below) uses discretized area, distance and centrality.

Feature	PRECEDE	Z	PREC.-EST.	Z	INTER	Z	FOLLOW	Z
intercept	-4.18 ± .37	-11.2	-2.66 ± .50	-5.3	-2.51 ± .32	-7.7	2.72 ± .32	8.5
anch area	-.27 ± .06	-4.6	-.19 ± .09	-2.2	-	-	.35 ± .05	6.9
anch centr	.11 ± .05	2.0	-	-	-	-	-	-
anch deps	-	-	-.74 ± .12	-6.2	.22 ± .06	3.6	-	-
anch=targ	.30 ± .13	2.3	-	-	.55 ± .14	4.0	-.71 ± .13	-5.7
distance	-	-	-.24 ± .09	-2.6	-	-	-	-
lmk=reg	11.46 ± 1.35	8.5	-	-	3.01 ± 1.19	2.5	-12.62 ± 1.17	-10.8
lmk area	3.27 ± .38	8.7	-	-	1.28 ± .32	4.0	-3.76 ± .32	-11.7
lmk centr	-	-	-	-	-	-	.81 ± .32	2.6
lmk deps	-	-	2.38 ± .14	16.9	-1.07 ± .13	-8.3	-1.37 ± .12	-11.5

Table 2: Regression coefficients, standard deviations and Z-scores from one-vs-all logistic regressions with direction/ESTABLISH status as output variable. Only effects significant at $p < .05$ level are shown; other effects are displayed as -.

alized at the start of an expression. (Clarke et al. (2013) show that they have fewer landmarks overall.) Again, the effect of centrality is counterintuitive, but weak ($\beta = .81$). Anchors with more dependents are slightly more likely to use the INTER slot ($\beta = .22$), suggesting that the various dependents are spread syntactically throughout the expression.

Although distance and centrality are weak indicators in this dataset, area shows strong effects which support our conclusion that visual salience behaves like discourse salience. The standard information order of English clauses places given information first and new information later (Prince, 1981). Thus, we observe that the non-right orders are used for larger objects, which is what we would expect if their visual perceptibility is sufficient to place them in common ground despite the lack of a previous mention.[10] On the other hand, the FOLLOW order is used for smaller objects that cannot be assumed to be part of common ground (and are therefore treated as new).

The use of ESTABLISH constructions for mid-sized objects also makes sense on theoretical grounds. ESTABLISH constructions are a way of achieving the PRECEDE information structure, which places the landmark first— and this makes sense primarily if the landmark is reasonably salient, since otherwise it will not be found any faster than the target. On the other hand, most of the constructions we discuss as ESTABLISH, such as existential "there is", require their object to be discourse-new (Ward and Birner, 1995); it would be infelicitous to start a description by stating the existence of something already in the common ground "there is a sky, and it is blue…" Thus, it makes sense that neither large or small objects favor the use of this construction; it can be used to foreground an object which is not salient enough to be assumed in common ground, but *is* salient enough to find without a great deal of visual search.

8 Information structure prediction

In this section, we experiment with an idealized version of the information structuring task. We provide our system with gold standard content selection— we know which objects will be mentioned, and if they serve as landmarks, we know the anchor they describe. However, we do not know which information strategies will be used to order them; our task is to predict this. In doing so, we are working with an idealized version of the standard generation pipeline, which often operates as a two-stage process, with content selection followed by surface realization. Information structure prediction is intermediate between these two stages; once we have decided which objects to mention (or in concert), we would like to decide what order to mention them in.

We set up the prediction task as in the previous section: Given an anchor/landmark pair, our system must decide what direction and ESTABLISH status to assign it. However, here we evaluate the system as a classifier. We treat anchor/landmark pair as independent from the others

[10] Prince (1981) discusses other discourse-new items that are nonetheless treated as familiar, like "The FBI", under the name *unused* (that is, available, but not previously in use in the discourse).

Feat type	# features
type (targ/lmark/region) of anchor	3
type (targ/lmark/region) of dep	3
quartile of anchor area	4
quartile of lmark area	4
quartile of anchor → lmark dist	4
quartile of dist anchor → screen ctr	4
quartile of dist lmark → screen ctr	4
# direct dependents of anchor	6
# descendents of anchor	6

Table 3: Feature templates and number of instantiations in our discriminative system.

(including other pairs from the same description); during development, we investigated a parser-like structured classifier based on (Socher et al., 2011; Salakhutdinov and Hinton, 2009) that jointly classified all the relational descriptions in a single utterance at once, but results did not improve over the classifier system, perhaps because on average the trees are fairly shallow.

8.1 Discriminative comparison

We train a discriminative multilabel classifier using maximum entropy.[11] We predict EST-DIR pairs given a set of discrete features shown in Table 3. This setup differs slightly from the previous section (which used one-vs-all); we are attempting to conform to the standard practices of psycholinguistics and computational linguistics respectively. Area, salience, distance to center and inter-object distance values are discretized by determining in which quartile of the training set each value falls (lowest 25%, mid-low, mid-high, highest 25%). Our initial model used continuous values as in the previous section, but results were somewhat poorer, suggesting some of these features may have nonlinear effects.

8.2 Experiments

We hold out three images (*vikings*, *airport*, *blackandwhite*) as a development set. In test, we exclude these 3 documents and use the other 8 for evaluation. In both development and test, we conduct experiments by crossvalidation, testing on one document at a time and training on the other ten.[12]

We report two trivial baseline strategies, all landmarks following (the best baseline for overall accuracy) and all landmarks preceding (the best baseline for predicting the direction, but not as good overall because the PRECEDE predictions are split between ESTABLISH and not ESTABLISH). Our preliminary analysis shows that regions have a strong tendency to precede their anchors, so we also report results for a baseline using this pattern (regions preceding, everything else following). We believe this baseline pattern is the one which would be learned as a template by previous systems like Di Fabbrizio et al. (2008), since this system can learn relationships between broad types of entities (target, landmark and region) but does not use visual features of the actual entities in the scene to make any finer distinctions.

We also provide two "inter-subject" oracle scores intended to estimate the performance ceiling imposed by human variability. This oracle assigns each anchor/landmark pair the direction and ESTABLISH status assigned by the majority of speakers who mentioned that pair. The "multiple mentions" estimate of agreement is the one mentioned in Section 5; it was based only on pairs mentioned by multiple speakers. The "all" estimate is based on all objects; it is higher because, for pairs mentioned by only one speaker, it is by definition perfect. Our system's use of the number of descendants feature is not captured by this oracle— these features capture information about a particular speaker's content plan beyond their decision to mention a particular pair— but we suspect that the oracle's performance will nonetheless be hard for any practical system to beat.

We report gross accuracy (correctly predicting both DIR and ESTABLISH) for relational pairs (Table 5), and also decompose by direction (Table 4) and ESTABLISH status (Table 6).

The baseline correctly predicts 43% of pairs, implying that this pattern (regions precede, landmarks follow) covers a bit under half the data. The classifier improves this to 52%. When predicting the direction alone, the best baseline (PRECEDE) scores 42%; the classifier scores 57%. All system scores are significantly better than the baseline (sign test on pairs, $p < 0.01$). In predictions of ESTABLISH tags, our result is a 60% f-score, which is indistinguishable from the lower bound

[11] Learned using the Theano neural-network package (Bergstra et al., 2010) and stochastic gradient descent code from `deeplearning.net/tutorial` (Bengio, 2009).

[12] This means we always use 10 of the 11 documents for training, whether in dev or test, but we didn't do error analysis on the training examples. Data size does appear to matter; training on 8 documents at a time and testing on 3 yields poorer results.

System	PRECEDE			INTER			FOLLOW			Dir Acc
	Prec	Rec	F	Prec	Rec	F	Prec	Rec	F	
Follow	0	0	0	0	0	0	32	100	49	32
Precede	44	100	62	0	0	0	0	0	0	44
Regions precede	61	32	42	0	0	0	37	87	52	42
Discr	66	69	68	39	23	29	53	65	58	57
Inter-subj (multiple mentions)	77	61	68	54	62	58	67	76	71	66
Inter-subj (all)	84	75	79	65	69	67	74	83	78	76

Table 4: Direction scores (p/r/f per direction and total pair directions correctly predicted) in 2382 pairs in test set. Overall accuracy differences between system and baselines are significant ($p < .01$).

System	Pair accuracy
Follow	36
Precede	29
Regions precede	43
Discr	52
Inter-subj (mult)	64
Inter-subj (all)	74

Table 5: Gross accuracy (%) for 2382 test pairs.

System	ESTABLISH		
	Prec	Rec	F
Follow	0	0	0
Precede	0	0	0
Regions precede	0	0	0
Discr	55	67	60
Inter-subj (mult)	68	43	53
Inter-subj (all)	82	66	73

Table 6: ESTABLISH scores (p/r/f for EST=TRUE) in 2382 pairs in test set.

estimate of interannotator agreement.

9 Conclusions

The results of this study show that the information structure of relational descriptions is highly variable, and depends on notions of salience and common ground that are difficult to capture with templates or simple case-based rules. This suggests that the question of realization for visual-word referring expressions may need to be reopened. A data-driven approach not only allows better prediction of which strategy will be used (reducing error by 9% absolute, 16% relative) but also enables us to analyze the pattern and conclude that the visual salience of an object acts in the same way as discourse salience.

Several open questions remain. One is the failure of the Torralba et al. (2006) visual distinctiveness model to make any difference: Is this actually a perceptual fact, or does it merely demonstrate that the model is not as predictive of human attentional patterns as we would like? More important is the question of what lies behind the substantial variations we observe across individuals. These may reflect truly different strategies; for instance, some speakers may generate REs incrementally as they scan the image (Pechmann, 2009) while others perform a more complete scan before beginning (Gatt et al., 2012). We suspect answering this question is beyond the scope of corpus studies, and intend to investigate via psycholinguistic experiments using an eyetracker.

Another question is to what extend the patterns we observe are intended to facilitate listeners' visual search (an audience design hypothesis) versus speakers' efficient construction of utterances. This study focused on predicting speaker behavior, while acknowledging that the utterances speakers produce are not always optimal for listeners (Belz and Gatt, 2008). However, we suspect that in this case, putting easy-to-see objects early really does help listeners; we are currently planning perception experiments to test this hypothesis.

Finally, we intend to incorporate the visual features used in this study into a full-scale realization system. This will enable us to create more human-like REs for visual domains. Such REs can be incorporated into natural language systems for a variety of interactive visual-world tasks.

Acknowledgements

The third author was supported by EPSRC grant EP/H050442/1 and ERC grant 203427 "Synchronous Linguistic and Visual Processing". We also thank Marie-Catherine de Marneffe, Craige Roberts, the OSU Pragmatics group and our anonymous reviewers for their helpful comments.

References

D. A. Baldwin. 1995. Understanding the link between joint attention and language. In *Joint attention: its origins and role in development*. Lawrence Erlbaum Assoc., Hillsdale, NJ.

D. Bates, M. Maechler, and B. Bolker. 2011. lme4: Linear mixed-effects models using s4 classes. Comprehensive R Archive Network: `cran.r-project.org`.

David L. Bean and Ellen Riloff. 1999. Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL'99)*, pages 373–380, Morristown, NJ, USA. Association for Computational Linguistics.

Anja Belz and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 197–200. Association for Computational Linguistics.

Yoshua Bengio. 2009. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127. Also published as a book. Now Publishers, 2009.

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June. Oral Presentation.

Robbert-Jan Beun and Anita H.M. Cremers. 1998. Object reference in a shared domain of conversation. *Pragmatics and Cognition*, 6(1-2):121–152.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. *Cognitive Foundations of Interpretation*, pages 69–94.

Ivo Brugman, Mariët Theune, Emiel Krahmer, and Jette Viethen. 2009. Realizing the costs: template-based surface realisation in the graph approach to referring expression generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 183–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Carpenter, K. Nagell, and M. Tomasello. 1998. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, 63(4).

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge.

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.

Michael Crawley. 2007. *The R Book*. Wiley-Blackwell, Hoboken, NJ.

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

Giuseppe Di Fabbrizio, Amanda J. Stent, and Srinivas Bangalore. 2008. Referring expression generation using speaker-based attribute selection and trainable realization (ATTR). In *Proceedings of the 5th International Conference on Natural Language Generation (INLG)*, Salt Fork, OH.

Manjuan Duan, Micha Elsner, and Marie-Catherine de Marneffe. 2013. Visual and linguistic predictors for the definiteness of referring expressions. In *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*, Amsterdam.

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

Rui Fang, Changsong Liu, Lanbo She, and Joyce Y. Chai. 2013. Towards situated dialogue: Revisiting referring expression generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 392–402, Seattle, Washington, USA, October. Association for Computational Linguistics.

Katja Filippova and Michael Strube. 2007. Generating constituent order in German clauses. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 320–327, Prague, Czech Republic, June. Association for Computational Linguistics.

William Ford and David Olson. 1975. The elaboration of the noun phrase in children's description of objects. *Journal of Experimental Child Psychology*, 19:371–382.

Albert Gatt and Anja Belz. 2010. Introducing shared task evaluation to NLG: The TUNA shared task evaluation challenges. In E. Krahmer and M. Theune, editors, *Empirical Methods in Natural Language Generation*. Springer, Berlin and Heidelberg.

Albert Gatt, Anja Belz, and Eric Kow. 2008. The TUNA-REG challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Conference on Natural Language Generation (INLG)*, Salt Fork, OH.

Albert Gatt, Anja Belz, and Eric Kow. 2009. The TUNA-REG challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, Athens.

A. Gatt, E. Krahmer, R. P. G. van Gompel, and K. van Deemter. 2012. Does domain size impact speech onset time during reference production? In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, pages 1584–1589, Sapporo, Japan.

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

M. Handford. 1987. *Where's Wally?* Walker Books, London, 3 edition.

M. Handford. 1988. *Where's Wally Now?* Walker Books, London, 4 edition.

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

John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *ACL*.

J. Kelleher, F. Costello, and J. van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167(12):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.

Claudia Maienborn. 2001. On the position and interpretation of locative modifiers. *Natural Language Semantics*, 9(2):191–240.

Crystal Nakatsu and Michael White. 2010. Generating with discourse combinatory categorial grammar. *Linguistic Issues in Language Technology*, 4(1).

T. Pechmann. 2009. Incremental speech production and referential overspecification. *Linguistics*, 27(1):89–110.

Ellen Prince. 1981. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York.

Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Replicated softmax: an undirected topic model. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1607–1614.

Richard Socher, Cliff C. Lin, Andrew Y. Ng, and Christopher D. Manning. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*.

A. 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.

A. Torralba, A. Oliva, M. Castelhano, and J. 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 and Robert Dale. 2008. The use of spatial relations in referring expressions. In *Proceedings of the 5th International Conference on Natural Language Generation*, Salt Fork, Ohio, USA.

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

Gregory Ward and Betty Birner. 2001. Discourse and information structure. In Deborah Schiffrin, Deborah Tannen, and Heidi Hamilton, editors, *Handbook of discourse analysis*, pages 119–137. Basil Blackwell, Oxford.

Bonnie L. Webber. 2004. D-ltag: extending lexicalized tag to discourse. *Cognitive Science*, 28(5):751–779.

Michael White and Rajakrishnan Rajkumar. 2012. Minimal dependency length in realization ranking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 244–255, Jeju Island, Korea, July. Association for Computational Linguistics.

Jeremy M. Wolfe. 2012. Visual search. In P. Todd, T. Holls, and T. Robbins, editors, *Cognitive Search: Evolution, Algorithms and the Brain*, pages 159 – 175. MIT Press, Cambridge, MA, USA.

Sina Zarrieß, Aoife Cahill, and Jonas Kuhn. 2012. To what extent does sentence-internal realisation reflect discourse context? a study on word order. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 767–776, Avignon, France, April. Association for Computational Linguistics.