OMEGA: A Probabilistic Approach to Referring Expression Generation in a Virtual Environment

Maurice Langner
Department of Linguistics
Ruhr-Universität Bochum
Maurice.Langner@rub.de

Abstract

In recent years, referring expression generation algorithms were inspired by game theory and probability theory. In this paper, an algorithm is designed for the generation of referring expressions (REG) that base on both models by integrating maximization of utilities into the content determination process. It implements cognitive models for assessing visual salience of objects and additional features. In order to evaluate the algorithm properly and validate the applicability of existing models and evaluative information criteria, both, production and comprehension studies, are conducted using a complex domain of objects, providing new directions of approaching the evaluation of REG algorithms.

1 Introduction

Probabilistic and game-theoretic approaches to REG base on the maximization of utilities (Frank and Goodman, 2012; Goodman and Frank, 2016) or on corpus frequencies of attributes that are selected according to random float values (e.g. the PRO model, Gompel et al., 2019). These models and their humanlikeness of production were tested on domains of limited size, mostly three objects, and a minimal set of 3 to 4 properties. The parameters of these models are learned using Maximum Likelihood Estimation. Due to the non-deterministic nature of probabilistic models, evaluation metrics used for deterministic models, especially the Dice Score, cannot be applied. Therefore, the likelihood of the predicted distribution, given the observed data, is computed using the Bayesian Information Criterion (BIC). With a limited set of possible referring expressions, the likelihood function, which bases on the multinomial distribution, produces reasonable results, but an increasing complexity of the domain may necessitate a different evaluation approach.

Empirical data against which the reference games model (Frank and Goodman, 2012; Franke and Jäger, 2016) was tested has been collected in betting games in a forced-choice paradigm. Such data sets are not only non-reproducible for larger domains, but also circumvent the integration of underlying cognitive processes of content determination into the model. Hence, the aim of this paper is to develop a new probabilistic REG-algorithm that is able to handle the choice of complex attributes and spatial relations on the basis of a cognitively motivated tradeoff between salience, preference and discriminatory power (DP). An important aspect of this tradeoff must be the possibility of generating minimal and overspecified expressions in complex domains for overcoming the deficiency of models rooted in game theory, which typically produce unary expressions only. Overspecification is of essential importance for a humanlike modeling of reference, since overspecification has been proved to facilitate the identification task for the hearer if the redundant information is easily accessible (Paraboni et al., 2017).

In what follows, first the domain underlying the experimental studies will be presented. Then the production survey is described. Subsequently, cognitive models for the approximation of visual salience are introduced which are used in the OMEGA algorithm. Section 5 gives a detailed description of the developed model. Section 6 evaluates the algorithm performance in both humanlikeness and comprehension, giving insight into problems and advantages of deployed evaluation methods.

2 The reference domain

The domain for the empirical studies consists of 3D geometric primitives with different properties. Atomic properties are SHAPE and NAME TAG,

Proceedings of The 13th International Conference on Natural Language Generation, pages 296–305,
Dublin, Ireland, 15-18 December, 2020. ©2020 Association for Computational Linguistics

while SIZE is a scalar property whose value is determined in the context, and COLOUR relies on the calculation of shades for RED, BLUE and GREEN. The light and dark shades are each prefixed with the according specifier *light-* or *dark-*. Two further relational properties are given as coordinates for x-axis and y-axis. The design of this corpus is different from other well-renowned corpora (e.g. TUNA) in several important aspects. Colors provide different degrees of visual difference. Size is varying by fixed ratios. The domain on which the production data from participants is collected must agree with the domain for the comprehension study. Last but not least, the type attribute is fixed as *gift box*, such that the generation of the SHAPE attribute does not vary.

ATTRIBUTE	VALUES
colour	{(light-, dark-) blue, red, green, grey, black, brown}
type	{ball, cube, torus, cone, cylinder, flower}
size	{0.5, 0.75, 1}
tag	{circle, square}
x-coord.	{1,2,3,4,5}
y-coord.	{1,2}

Table 1: object domain: attributes and their values

Each object is rendered with each possible attribute combination in a 3D setting with identical lighting and orientation, which results in a domain with 360 different objects.

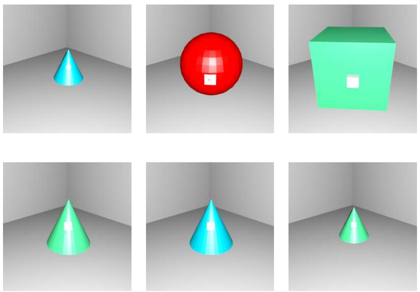


Figure 1: Six sample objects from the domain

In Figure 1, six sample objects are given. Assuming that the top left picture is the target, the referring expression *the cone* would be insufficient for identification; a minimal expression would contain expressions for COORDINATES or COLOUR, SIZE and TYPE. Adding further attributes to these combinations results in overspecification, e.g. *the small light-blue cone in the top row*.

3 Empirical production survey

A web-based empirical survey has been conducted in order to collect production data from native speakers of English. This data reveals the attribute choice and reference habits of the participants. The production survey is split into two subexperiments. In the first part, speakers are rating the visual salience of attribute values for each attribute separately, e.g. in the set of COLOUR values, participants rate the salience of each value within the set of colours. Additionally, for SIZE and COLOUR, the correlation between the gradience and the salience ratio is assessed by letting the participants rate the visual difference between values. An example of this assessment of the visual difference is given in Figure 2.

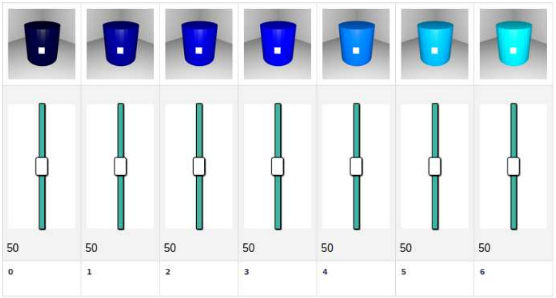


Figure 2: surface of the production survey, part I: rating the visual difference between colour gradients

The intention behind this survey design is to provide an attribute-specific rating of visual salience or prominence, respectively. These values serve as input to the REG models presented in the following sections. The downside of this is the dependence of models on corpora for each unseen domain. For this reason, section 4 proposes cognitive models that may replace corpus data and make REG models independent from corpora and collected production data. The second part of the study comprises production sessions. In each session, between 3 and 9 objects are depicted in a grid (see Figure 1). The speaker's task is to compose an expression that uniquely identifies the target object which is highlighted by a golden frame. Participants were able to select and deselect attribute values, candidate objects for spatial relations and in turn attributes for the description of the spatially related object. Attribute selection is restricted to the choice from domain attributes and their values in order to avoid the necessity of normalizing free text input after collecting the corpus data. Participants could also produce a surface realization from their selected attributes. An intriguing fact is that this restriction of attribute selection does not trigger a shift in preferential usage of attributes in comparison to e.g. the TUNA corpus and other psycholinguistic studies (e.g. Tarenskeen et al., 2015). The production subdomains of 10 sessions were fixed, 15 subsequent sessions were randomized. The web application allowed for the integration of identical objects several times, such that identification may be possible only via using coordinates. Via Prolific, 100 native speakers of English from the United Kingdom were acquired. The produced corpus consists of about 2500 production sessions. The pie chart in figure 3 reveals the amount of overspecified phrases with 60% of all produced expressions in the data. The capability of producing overspecified expressions is consequently of high importance for humanlikeness in the REG task. The outer circle is divided into expressions with and without spatial relations, while the inner circle visualizes the sum of the respective generation quality.

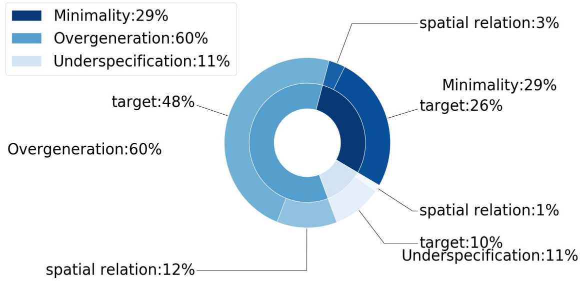


Figure 3: Generation qualities of the production corpus

4 Cognitive models for approximation of visual salience

A well justified metric is needed in order to model the perceptual prominence of each object as a substitute for salience ratings. If this salience is a joint effect of the visual salience of the object's properties, a utility function for a REG model needs to integrate a computation of how much the object springs to the speaker's eye. For size, this is straightforward, since the domain objects were rendered with three different size factors. The prior production survey reveals a linear relation between the increase in size factor and the increase in perceptual size difference. Van Deemter and his co-authors (2016, p.141; van Gompel et al., 2014) prove that in REG, the ratio of size difference has essential importance on the production probability of this property. Therefore, for a REG model, this ratio models the salience for size in a set of objects. For COLOUR, comparing values is more complex. Cognitive experiments on colour perception led to the conception of the CIElab colour space (Mokrzycki and Tatol, 2011). Formulas have been construed that calculate the difference between two colours, the ΔE value, that humans may still perceive. Multiple revisions of the CIElab ΔE functions exist with different terms for compensating hue and saturation, but they all base on the Euclidean distance of colours which are represented as 3-dimensional vectors in the CIElab colour space. These ΔE functions allow to compute the difference between two colours in a cognitively reasonable way and can be used in a utility function.

A challenge to be faced is that in realistic settings corresponding metrics are required for other attributes as well, e.g. shapes. Although not implemented in the model presented in this paper, attempts are made to formalize the similarity measures between shapes by analysing their vertices (Mori et al., 2001) and points on the contour (Ling and Jacobs, 2007). Latecki and Lakaemper (2000) developed a model for shape comparison that is close to cognitive processes of noisy shape perception. Relational attributes may be modeled using an approach comparable to Kelleher and Genabith's (2004) measure of assessing the distance towards the center of the visual field, but this is subject to further research. The main advantage of approximating visual salience with these cognitive models is that values are calculated in situ on the present visual field, instead of optimizing parameters such that they fit a large data set best. In this way, referring expressions may be generated in a more individual way and be tailored to the present domain, while other methods rather capture the whole data on average. Finally, these cognitive models make REG models independent from experimental studies as given in section 3 under the premise that they perform equally well.

5 The OMEGA model

The overspecifying, utility maximizing referring expression generation algorithm, short OMEGA, is an incremental, non-deterministic and probabilistic algorithm which uses a utility function in order to calculate which attributes are the most utile ones for production. This includes the choice of a candidate object for spatial relations and the attribute selection for the description of this candidate.

At its core, a utility function as used in reference games (Frank and Goodman, 2012) computes production probabilities which guide content determination. The concept of this utility function is crucial for the performance of the whole model. At this point, an observation made on the empirical data from the prior study becomes a decisive factor.

When high DP coincides with high salience, the usage of the respective attribute increases drastically, while high DP for non-salient attributes does not influence attribute selection significantly. Additionally, van Deemter (2016) and van Gompel et al. (2014) show that larger proportional differences in gradable attributes, e.g. SIZE, trigger a higher usage in referring expressions than small differences that are cognitively less easily assessable. After all, this is the interference between DP and salience. The 'object-distractor contrast', as Mast (2016, p.140) terms it, in regard to a specific attribute is influenced both by the number of distractors that provide the same attribute value as well as by the visual distinctiveness of the attribute value in comparison with other values of the same attribute that apply to the distractors. Therefore, salience, preference and DP must be taken balanced when defining the utility function.

The salience computation which is used in the utility definition of the OMEGA algorithm integrates the cognitive models of perception presented in the previous section. The formulae below describe how the salience values are computed. The attribute value of the target object is compared to the values of the distractors in regard to that attribute and the average value is calculated. Therefore, salience is represented as the average visual difference between the target and the set of distractors D. The more different, the higher is the salience. The factor α projects the Δ_E value onto the interval of numbers between 0 and 1, c is a colour vector.

$$SAL_{colour}(c, D) = \frac{\sum_i (\alpha \Delta_E(c, c_i))}{|D|} \tag{1}$$

For size, the absolute value of the difference between s and s_i is needed in order to prevent negative values when $s < s_i$.

$$SAL_{size}(s, D) = \frac{\sum_i abs(s - s_i)}{|D|} \tag{2}$$

The same method is used for discriminatory power which is a ratio of how many distractors an attribute value rules out. Considering the fact that attribute selection is performed across attributes (and not across attribute values), it is possible to calculate for each attribute, of which a value applies to the target, what the average DP is in comparison with values of the respective attribute that apply to the distractors.

$$AVG_{DP}(e, D) = \frac{\sum_e max(0, DP(e, D) - DP(e_i, D))}{|D|} \tag{3}$$

In order to prevent negative utilities, the function only adds the difference of e to some $e_i \in D$ to the sum if the result is > 0. In other words, if the DP of the target attribute is smaller than the DP of a competitive attribute, the negative distance is not added, but the average DP nevertheless is decreased since it is normalized against the number of all competitive attributes, i.e. including those with higher DP. For the OMEGA model, the utility function u is defined as the product of the average DP and the salience of the value v_A of attribute A that is true of the target object in domain D. To this product, the preference degree estimated by the frequency of attributes in the production data can be added under the assumption that preference is not independent from the joint effect of salience and DP. The decision to integrate preference is also motivated by the findings of Ferreira and Paraboni (2014), who prove that the integration of speaker preferences enhances the quality of their SVM classifiers for REG.

$$\begin{aligned} U(v_A, A, D) = AVG_{DP}(v_A, D) \\ *SAL_A(v_A, D) + Pref(A) \end{aligned} \tag{4}$$

In RSA and other log-linear models a logarithm is used in order to map large utility values to a smaller interval of results (Qing and Franke, 2015; Franke and Degen, 2016). This has the positive effect of assigning the lowest possible utility, namely negative infinity, to attributes with a literal interpretation of 0, or, as in OMEGA, irrelevant information. In combination with softmaximizing utilities, the model conforms with Grice's Maxim of Quantity. Softmaximization which is deployed in OMEGA means turning the utilities into weighted production probabilities, as shown in the formula (5).

$$P(e|r, D) = \frac{exp(\lambda Utility(e; r, D))}{\sum_{e_i \in D} exp(\lambda Utility(e_i; r, D))} \tag{5}$$

By means of the chain rule we are able to calculate the probabilities of all candidate referring expressions e which are sequences of attribute values

a that are true of the referent. After attribute value a_1 is selected, the set of distractors D of domain M is updated, as well as the attribute a_1 belongs to is removed from the set of selectable properties. This probability is multiplied with the result of calculating the probability of a_2 given the updated data structures and the already selected attributes, and so on until a_n has been achieved. This mechanism permits the OMEGA model to produce a probability distribution over all possible expressions that are true of the target.

A closer analysis of the pseudo code in algorithm 1 will show how OMEGA implements the selection of reference objects and production of spatial relations, as well as the enrichment with redundant attributes for overspecification.

Line 4 initializes two empty data structures e and e_{rel} for attributes that are selected during the production process. In line 5, rel_c is a container for storing possible candidates from the set of distractors that may be used for a spatial relation. The subsequent while-loop (lines 6 to 14) calculates for each attribute a in the set of attributes A what production probability it provides in the given visual field. In line 9, the production probability is calculated, which integrates the utility function described earlier as well as the cognitive models for salience approximation. If the probability is higher than the one of the previous attribute, the maximally probable attribute is replaced with a. Then, data structures are updated. If some distractor is found for which the production probability for the present attribute a is larger then for the target, it is stored as a candidate for spatial relations (line 10). The while-loop terminates when either the set of distractors M or the set of attributes A is empty. In line 15, if some candidate for spatial relations has been found, OMEGA is recursively called on that object. Otherwise, the overspecification process starts (line 18). While overspecification is selected and A is not empty, maximal probable attributes are added to the expression using the same mechanism as in the first while loop.

Algorithm 1 Overspecifying Utility-Maximizing Referring Expression Generation Algorithm (OMEGA)

```
 1: Input: reference domain $D$ with target object $r$ and a
    non-empty set of distractors $M$, a set of attributes $A$ at
    least one of whose values is true of $r$.
 2: Output: the referring expression with maximal probability
 3: function OMEGA(r, D, M, A, Ω)
 4:     $e, e_{rel} \leftarrow [\,]$
 5:     $rel_c \leftarrow [\,]$
 6:     while $A \neq \emptyset \wedge M \neq \emptyset$ do
 7:         $a_{max} \leftarrow$ NONE
 8:         for $a \in A$ do
 9:             $p \leftarrow P(a|r, D)$
10:             $rel_c \leftarrow rel_c \cup \{r'|r' \in M \wedge P(a|r',D) > p\}$
11:             if $p > a_{max}$ then
12:                 $a_{max} \leftarrow a$
13:         $e.insert(a)$
14:         $update\ M, A$
15:     if $rel_c \neq \emptyset$ then
16:         $r_{rel} \leftarrow$ ARGMAX$(rel_c)$
17:         $e_{rel} \leftarrow$ OMEGA$(r_{rel}, D, M, A_{rel}, \Omega)$
18:     while RANDOMBOOLEAN$(\Omega)$ $is\ True \wedge A \neq \emptyset$ do
19:         $a'_{max} \leftarrow$ NONE
20:         for $a' \in A'$ do
21:             $p' \leftarrow P(a'|r, D)$
22:             if $p' > a'_{max}$ then
23:                 $a'_{max} \leftarrow a'$
24:         $e.insert(a')$
25:         $update\ A'$
26:     $return\ e, e_{rel}$
```

A possible point of criticism is that the usage of the formula for production probability may not be suitable for overspecification, since the utility tries to calculate the average salience and DP values on an empty set of distractors. In full awareness of this circumstance the functions for salience and DP are intended to integrate default values. For DP, the default value is 1, while for salience, the prior ratings are used. In summary this means, that the overspecification process is guided exceptionally by salience and preference, which is comparable to PRO and IA (Dale and Reiter, 1995; Reiter and Dale, 1992), while also being in perfect agreement with psycholinguistic studies on overspecification in REG (Tarenskeen et al., 2015). A legitimate criticism is indeed, that OMEGA does not cover the cases correctly where both target description and relational description are underspecified, but producing a minimal expression in joint effect. A modification to OMEGA that allows to cover this small number of expressions found in the empirical data remains open for further research. An intriguing point is that both selection steps share the same set of attributes, meaning that for overspecification the residual attributes are used that have not been selected beforehand. As a result, attributes that were not produced in the first step, for reasons of an average DP value close or equal to zero, are reconsidered in the overspecification step. Nonetheless OMEGA makes a giant leap towards independence from corpus data by implementing the cognitive models of salience approximation as a substitute for corpus statistics. The evaluation in the next section gives insight in which way the cognitive models influence the performance of the model in comparison to its competitors. Table 2 lists the technical differences between the competitor models and OMEGA.

feature	OMEGA	PRO	IA
parameters	1	|A|+1 where A is the set of attributes	0
decision model	prob.	prob.	determ.
spatial descriptions	yes	no	no
training	corpus statistics	optimization & MLE	corpus statistics
cognitive measures	yes	no	no
DP	yes	no	no
Salience	yes	no	no
Preference	yes	yes	yes

Table 2: Comparison of the competitor models

On the domain displayed in figure 1, these algorithms may generate different expressions. IA generates deterministically according to the preference order until all distractors are ruled out. This may also trigger overspecification. For the top left object as target, IA will add attributes according to preference degree, namely first COLOUR and TYPE, then SIZE. Since this is not discriminating yet, it further adds NAME TAG, resulting in a minimal expression *the small light-blue cone with a circular name tag*'. Since IA is deterministic, this expression is generated at every instance. PRO and OMEGA produce a probability distribution over all possible expressions. PRO offers optimized parameters for each attribute plus the overspecification parameter, which are fine-tuned to match a full data set. In this case, the visual difference between the light-blue and light-green objects is rather unsalient, except for the red sphere the visual field is rather unichrome. While PRO uses the preference degree optimized for the full data set in order to probabilistically choose to (or not to) add colour, OMEGA recognizes the small visual salience on the domain in situ and assigns to colour a low production probability accordingly. Since DP influences the selection as well, highly discriminating attributes, e.g. coordinates, are considered. COLOUR and TYPE are being reconsidered only for the overspecification process. Additionally, the visual prominence of the NAME TAG attribute with the value $circular$ for the target object is diminishingly small, although it has highest DP. PRO may consider this attribute as well, while OMEGA identifies the low perceptive recognizability, resulting in a tiny salience value, which decreases the production probability. Since OMEGA recognizes the comparably far more salient object, “the red sphere” with both unique and highly salient attribute values for COLOUR and TYPE, OMEGA will add a spatial description to the expression, resulting in, e.g., *the leftmost light-blue cone to the left of the red sphere*.

6 Evaluation

For evaluation, the OMEGA model has been tested competitively against the PRO model and the standard IA algorithm. Both, humanlikeness in production and performance in a comprehension task, have been evaluated.

6.1 Humanlikeness

For the production side, the Dice Score is not applicable due to the non-deterministic nature of both OMEGA and PRO. For this reason, often BIC is used to evaluate the performance of the models given the production corpus that is collected in the empirical production survey (van Deemter, 2016). For toy domains with only two or three attributes, the number of possible referring expressions is limited to 3 and 9, respectively. A probability distribution over those categories is therefore less noisy than in a domain with 6 attributes, which give $2^6 = 64$ combinations of attributes. Due to the number of categories, of which many did not occur in the corpus while being generated by both algorithms and vice versa, the Maximum-Likelihood-Method (MLE) failed for both models. The reason might be that the optimizer mainly modeled noise due to data sparseness of the empirical data. Therefore, the likelihood of the predicted distributions, given the corpus data, could not be assessed. Instead of training model parameters, corpus frequencies were used as model parameters for PRO, but still the likelihood was 0. A fact that may explain this conversion towards 0 is that even the probabilities that are calculated from the data, given the same data, have a likelihood of only 19.245e^{-51}. Gompel et al. (2019) list their values for experiments in a scenario that equals the reference games by Franke and Jäger (2016). In this simple domain, the smallest likelihood listed is about $4.9 * 10^{-69}$, which indicates that the likelihood on large domains with far more categories might be too small to be represented by a float value. Dice Score is inapplicable but may serve as a blueprint for an evaluation method that is at least capable of approximating the performance of the models.

PRO and OMEGA are both models that predict a probability distribution over a discrete number of possible expressions. The empirical data provides about 2500 data points. A referring expression of participant q_i in session s_i is predicted by both models on session s_i with a specific probability. For each data point, it is possible to determine with what chance it is produced by the models given the domain and the referent. Instead of calculating how likely the distribution predicted by a model is, given the empirical data, which also entails the interpretation of the empirical data as the ‘truth’ (Lewandowsky and Farrell, 2011, p.181), one can at least calculate the average probability with which the models would generate the referring expressions produced by the participants.

$$P_{Dice} = \frac{\sum_{i=1}^{n} P_M(e_i|r_i, D_i)}{N} \tag{6}$$

Formula (6) captures the intuition behind this criterion, which could be circumscribed as a probabilistic Dice Score. The Model P_M assigns a probability to the expression e_i that is contained in data point i, given the corresponding referent r_i and the domain D_i. By dividing the sum of these probabilities by the number of data points N, the formula calculates the average probability model M assigns to the referring expressions that were produced by the participants of the study.

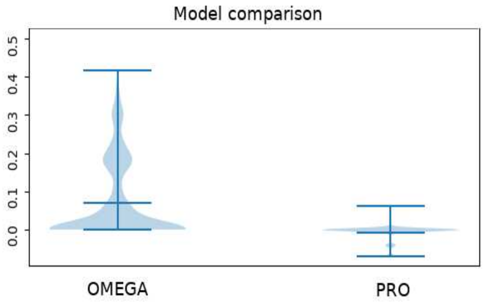
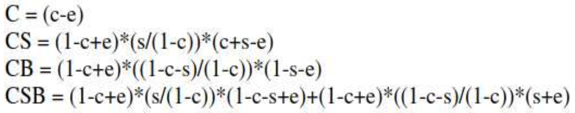


Figure 4: Probabilistic Dice Score

The according graph (figure 4) not only reveals that OMEGA performs better regarding this data set, but also that some probabilities that PRO assigns to expressions are negative, which means that some calculations in the PRO model cause some of the probabilities to get a negative polarity.

The reason for this is the optimization of parameters as probabilities for attributes, as well as the mathematical operations including overspecification eagerness. In figure (5), the probability of producing a unary expression COLOUR is $(c - e)$, where c is the probability that is proportional to the preference degree of COLOUR, whereas e is the parameter of overspecification eagerness. After optimization, the parameter for a rather unsalient attribute may be optimized in such a way that $c < e$, which causes $c - e$ to be negative. Effectively, the choice of subtraction or addition of a value e, that is not a probability by definition, from or to a probability, is not only theoretically questionable, but also leads to values that cannot be interpreted as probabilities. Indeed, data sparseness may have caused the optimizer to produce parameters that provided this configuration of $c < e$, but the problem is definitely rooted in the subtraction and addition of the probabilities that are assigned to the nodes of the decision tree. A possible remedy is using multiplication instead, such that the model is sound in the context of probability theory. Unfortunately, this has the downside of loosing both the elegant difference in mathematically combining overspecification eagerness with probabilities for continued overspecification and termination of the selection process.



C = (c-e)
CS = (1-c+e)*(s/(1-c))*(c+s-e)
CB = (1-c+e)*((1-c-s)/(1-c))*(1-s-e)
CSB = (1-c+e)*(s/(1-c))*(1-c-s+e)+(1-c+e)*((1-c-s)/(1-c))*(s+e)

Figure 5: PRO paths for a small domain with colour (C), size (S) and border (B), where colour is discriminating

Nevertheless, the modification did not significantly change the performance of PRO for humanlikeness. As described above, the evaluation methods for REG algorithms on the basis of empirical production data cannot be straightforwardly applied to probabilistic, non-deterministic models and larger domains. For BIC, data sparseness is the main reason for failure, since MLE cannot produce reasonable parameter estimates if the likelihood function provides no gradient. Dice Score cannot be applied to non-deterministic algorithms, since the score varies as much as the output does. Nonetheless, instead of calculating the sequential identity of expressions (Dice), the average probability can be determined which the models assign to the referring expressions in the empirical data.

6.2 Comprehension

Humanlikeness alone does not help to fully evaluate the performance of REG models. Human speakers are not always optimal, e.g. due to error-prone speaking. Evaluating humanlikeness only would mean modeling humanlike reference without knowledge of how well the referring expressions are understood and serve the purpose of identifying the target.

Another empirical study is necessary in order to assess how well human speakers comprehend the REs the models produce. Objective measures, e.g. identification time and error rate, give insight which model produces more optimal expressions in regard to comprehension.

The setting of the comprehension study in which the OMEGA model and its competitors are evaluated is a sales dialogue in a store for custom gift packages. A sales agent, which is simulated by the REG models, guides the participant through the survey and instructs him or her to find storage units and gift packages contained in these stores. The listener's task is to find the storage and identify the target object. The participant can freely move around in the virtual environment and perceives the rooms from the first person view. After identification of ten storage units and ten objects, the survey ends. Throughout the sessions, the system chooses randomly from the set of REG models, but the ten object domains are identical to the 10 fixed sessions in the production session in section 3.

For production of referring expressions, the sales agent (the generation system respectively) has access to the participant's position in the room and his direction of view. In this way, the model can give precise directional expressions that are tailored to the configuration of the virtual room and the participant's movements. Figure 7 depicts a screenshot from the participant's perspective on domain *S1*, for which OMEGA generates the following expressions:

1. *"Please got to the storage to the right of the antique column at the wall in front of you."*
2. *"Yes, here it is. Now, please find the red cylinder-shaped gift box."*

The sales agent gives feedback for each identification. For storage units, the system recognizes the participant's proximity to the storage as well as the direction of view. Participants identify gift boxes in these storage units by clicking when the cursor hovers above the object of their choice. Then, the sales agent gives feedback and either directs to the next storage on success or repeats the expression on false identification.

In this implementation, the referring expressions and instructions that are generated by the REG models are given as audio input. This reduces identification onset due to different reading speed and prevents distraction from the visual stimuli that may occur if the REs are presented as text. Participants can ask for repetition of the expression. The simulated sales agent then utters the expression again, but updates the RE according to the new player's position and direction of view.

The virtual environment (see layout in Figure 6) is less complex than the setting used for the GIVE challenges in regard to the room configuration (Koller et al., 2009,2010; Byron et al., 2007,2009; Striegnitz et al., 2011), since the number of rooms is smaller and there are no separate floors the participants need to be guided through. This was intentional, because the focus is shifted to referring expressions and not to instructions as used in navigation systems.

The study was conducted on 36 participants with sufficient knowledge of English. The participants completed the study on a local computer with a mouse, headset or loudspeaker. Two participants were excluded. The resulting corpus consists of 680 completed identification tasks and the respective measures of identification time, error rate and repetition counts.

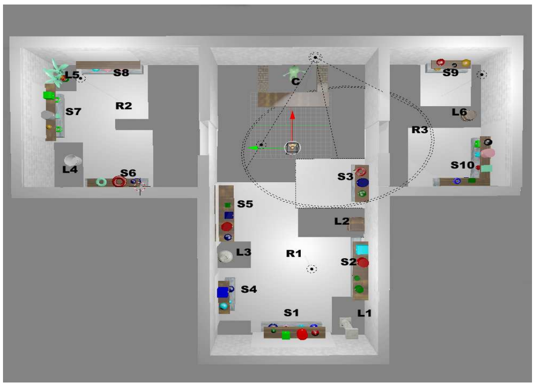


Figure 6: Layout of the virtual environment for the comprehension study

Average identification times are evaluated across participants for object identification and storage locating separately. The differences between PRO and OMEGA as well as IA and OMEGA are significant, with OMEGA having the highest average in identification times. The reason for this may be, that OMEGA is the only algorithm that is capable of producing spatial relations. In case that participants await the speaker to finish, this may increase the onset of the identification process. When spatial relations are excluded, the identification times for OMEGA with an average of 8.32 seconds are significantly lower than for PRO with an average of 10.62 seconds (F=4.006, p=0.04), while the difference between IA (average of 10.01 seconds) and OMEGA is close to significance (F=3.39, p=0.06). A straightforward conclusion to be drawn from this is that spatial relations prolong the identification process, while they may also increase proportionally with the length of the referring expression.

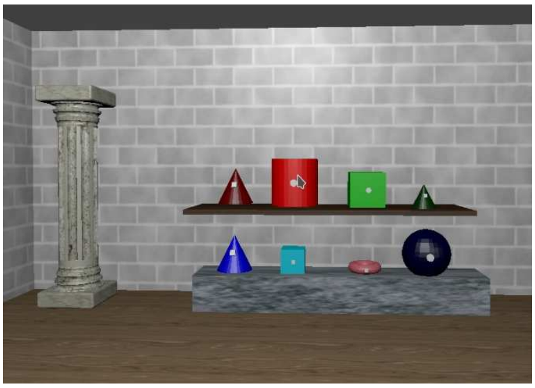


Figure 7: example domain from the virtual environment from the participant's perspective

For storage locating, there is no significant difference between PRO and OMEGA, but OMEGA is significantly faster than IA (F=3.89, p=0.049). An important observation to be made is that for IA on storage locating, the lowest times are more frequent than for PRO and OMEGA, but the distribution of times for IA is also much wider and outliers above 50 seconds are more frequent than for the competitors. A reason for this may be the generation quality, since IA tends far more often to produce minimal expressions, while PRO and OMEGA frequently overspecify.

In each comprehension session, during object identification, the application counted how often the participant selected a wrong object. This error rate is the second objective evaluation criterion to be used for comparing which of the three models performs better. Significance tests between data sets reveal that no difference between models is significant. Nonetheless, independent from the inclusion of spatial relations, OMEGA provides the smallest averages of errors with 0.87 and 0.85 respectively for expressions with and without spatial relations. IA is close to OMEGA with an average of 0.88 errors per session, while PRO is decently higher with 1.017 errors on average. Exclusion of spatial relations truncates the error rate at a maximum of three errors per session, but it also triggers a higher average. Since the difference is once more insignificant, no valid conclusion can be made, but the difference between the distributions indicates that spatial relations may prolong the identification process, but also show a tendency to reduce the error rate.

As one of the most important observations, a comprehension study for REG models in a virtual environment is subject to variance between participants as much as other empirical studies that collect language data from human speakers and listeners. Nevertheless, the comprehension study elicited a significant difference between OMEGA and its competitors in relation to the time that participants need to identify the target object. The data set indicates that OMEGA produces referring expressions that are more optimized to listener comprehension than PRO and IA, with a lower average error when spatial relations are given.

7 Conclusion

The OMEGA model succeeds in integrating the tradeoff between the most important parameters of REG, namely discriminatory power, salience and preference. The maximization of their joint effect, which defines the utility, allows to incrementally and probabilistically build referring expressions that may also contain spatial relations. The OMEGA model outperforms both IA and PRO in a comprehension task in a virtual environment. Nonetheless, more research is needed to further optimize the cognitive models that compute the salience values for different attributes which are needed for the OMEGA model. These measures provide the advantage of calculating salience in situ on the preset visual field, resulting in more independence from empirical input and parameter tuning at an improved performance level. One of the most important findings is the inapplicability of information criteria like BIC and optimization methods like MLE to models, given complex domains with numerous, realistically infinite, possible referring expressions. This is a reasonable indicator for the shift to comprehension as the standard evaluative method for REG models.

References

Donna Byron, Alexander Koller, Jon Oberlander, Laura Stoia, and Kristina Striegnitz. 2007. Generating instructions in virtual environments (GIVE): A challenge and an evaluation testbed for NLG. *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*.

Donna Byron, Kristina Striegnitz, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2009. Report on the first NLG challenge on generating instructions in virtual environments (GIVE). *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19.

Kees van Deemter. 2016. *Computational Models of Referring: A Study in Cognitive Science*.

Thiago Ferreira and Ivandré Paraboni. 2014. Referring expression generation: Taking speakers' preferences into account. *International Conference on Text, Speech and Dialogue*, 8655:539–546.

Michael Frank and Noah Goodman. 2012. Predicting pragmatic reasoning in language games. *Science (New York, N.Y.)*, 336:998.

Michael Franke and Judith Degen. 2016. Reasoning in reference games: Individual- vs. population-level probabilistic modeling. *PLOS ONE*, 11:e0154854.

Michael Franke and Gerhard Jäger. 2016. Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35.

Roger Gompel, Kees van Deemter, Albert Gatt, Rick Snoeren, and Emiel Krahmer. 2019. Conceptualization in reference production: Probabilistic modeling and experimental testing. *Psychological Review*, 126:345–373.

Roger van Gompel, Albert Gatt, Emiel Krahmer, and Kees van Deemter. 2014. Overspecification in reference: Modelling size contrast effects. *Proceedings of the 2014 conference on architectures and mechanisms in language processing*.

Noah Goodman and Michael Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20.

John Kelleher and Josef Genabith. 2004. Visual salience and reference resolution in simulated 3-d environments. *Artif. Intell. Rev.*, 21:253–267.

Alexander Koller, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, Jon Oberlander, and Kristina Striegnitz. 2009. The software architecture for the first challenge on generating instructions in virtual environments. *Proceedings of the Demonstrations Session at EACL 2009*, pages 33–36.

Alexander Koller, Kristina Striegnitz, Andrew Gargett, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2010. Report on the second nlg challenge on generating instructions in virtual environments (GIVE-2). *Proceedings of the 6th International Natural Language Generation Conference*.

Longin Jan Latecki and Rolf Lakaemper. 2000. Shape similarity measure based on correspondence of visual parts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22:1185 – 1190.

Stephan Lewandowsky and Simon Farrell. 2011. *Computational Modeling in Cognition: Principles and Practice*.

Haibin Ling and David Jacobs. 2007. Shape classification using the inner-distance. *IEEE transactions on pattern analysis and machine intelligence*, 29:286–99.

Vivien Mast. 2016. *Referring Expression Generation in Situated Interaction*. Ph.D. thesis.

Wojciech Mokrzycki and Maciej Tatol. 2011. Color difference delta E - A survey. *Machine Graphics and Vision*, 20:383–411.

G Mori, Serge Belongie, and J Malik. 2001. Shape contexts enable efficient retrieval of similar shapes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:I–723.

Ivandré Paraboni, Alex Lan, Matheus Ana, and Flávio Coutinho. 2017. Effects of cognitive effort on the resolution of overspecified descriptions. *Computational Linguistics*, 43:1–14.

Ciyang Qing and Michael Franke. 2015. Variations on a bayesian theme: Comparing bayesian models of referential reasoning. In Henk Zeevat and Hans-Christian Schmitz, editors, *Bayesian Natural Language Semantics and Pragmatics*, pages 201–220.

Ehud Reiter and Robert Dale. 1992. A fast algorithm for the generation of referring expressions. *Proceedings of the 15th International Conference on Computational Linguistics*, 1.

Kristina Striegnitz, Alexandre Denis, Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Mariet Theune. 2011. Report on the second second challenge on generating instructions in virtual environments (GIVE2.5). *Applied Geomatics*.

Sammie Tarenskeen, Mirjam Broersma, and Bart Geurts. 2015. Overspecification of color, pattern, and size: Salience, absoluteness, and consistency. *Frontiers in Psychology*, 6.