Language, Cognition and Neuroscience, 2014
Vol. 29, No. 8, 950–974, http://dx.doi.org/10.1080/01690965.2013.827224

Routledge
Taylor & Francis Group

Referring in dialogue: alignment or construction?

Jette Viethen[a]*, Robert Dale[b] and Markus Guhe[c]

[a]*Tilburg Center for Cognition and Communication (TiCC), School of Humanities, Tilburg University, PO Box 90153, 5000, LE Tilburg, The Netherlands;* [b]*Department of Computing, Macquarie University, North Ryde, NSW, Australia;* [c]*School of Informatics, University of Edinburgh, Edinburgh, UK*

(*Received 14 March 2012; final version received 15 July 2013*)

Human speakers generally find it easy to refer to entities in such a way that their hearers can determine who or what is being talked about. In an attempt to model this behaviour, researchers in computational linguistics have explored the development of algorithms that operate in a deliberate manner, choosing attributes of an intended referent on the basis of their ability to distinguish that entity from its distractors. Psycholinguistic models, on the other hand, suggest that speakers align their referring expressions at several linguistic levels with those used previously in the discourse. This implies more subconscious reuse, and less deliberate choice, than is found in computational models of referring expression generation. Which of these is a more accurate characterisation of what people do? Do both models capture aspects of human referring behaviour? In this paper, we use a machine-learning approach to explore these questions. In our first study, we examine how underlying factors of the psycholinguistic and the computational models impact on the production of reference in dialogue. In our second study, we explore the psychological validity of another crucial aspect of some computational approaches to reference production: their serial dependency characteristic, whereby attributes are included in a referring expression based on which other attributes have already been chosen. The results of both studies suggest that the assumptions underpinning computational algorithms do not play a large role in people's referring behaviour.

Keywords: reference production; alignment; computational modelling; machine learning

Almost every utterance that a person makes includes at least one referring expression whose function is to pick out some intended referent for the hearer. These references are invariably successful, and when they fail, repair is quickly achieved. How do people manage to successfully refer with such apparent ease?

This question has been explored by a number of different disciplines, but perhaps most intensely by psycholinguistics and computational linguistics. Within the psycholinguistics community, there is now a long tradition of work that explores how a dialogue participant's forms of reference are influenced by those previously used for a given entity in the same dialogue. Most prominently, this line of work has been discussed in terms of the notions of *alignment* (Pickering & Garrod, 2004) and *conceptual pacts* (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986). The basic idea is essentially very simple: whenever possible, people refer by reusing aspects of forms of reference that have been used before. Computational work on referring expression generation (REG), on the other hand, has generally taken a more constructive approach to the problem, seeing the task as being concerned with choosing those attributes of an intended referent that distinguish it from the other entities with which it might be confused (see Krahmer & van Deemter, 2012, for an overview).

A large body of empirical evidence exists for the alignment and conceptual pact models within psycholinguistics. At the same time, the considerations underlying computational approaches are intuitively very appealing: surely, the need to use attributes that distinguish the target referent from other distractor objects plays some role in the production of referring expressions? Using a corpus of over 20,000 referring expressions from task-oriented dialogues, this paper presents a machine-learning approach that makes it possible to compare models corresponding to the psycholinguistic and the computational approaches and to try out models that combine the two. We suspected that neither approach by itself would tell the full story, and that models combining the two approaches would achieve the best match with human referring behaviour. However, the results from our first study show that the constructive computational approach does not appear to contribute much to the way human speakers build referring expressions. These results expand on a preliminary study we conducted on subsequent references only (Viethen, Dale, & Guhe,

*Corresponding author. E-mail: h.a.e.viethen@uvt.nl

2011a, b), because they show that even initial references can be composed without the constructive elements of the computational approach.

Most algorithms for the generation of referring expressions choose the attributes to be used one by one, in a way that makes the inclusion of each attribute dependent on which attributes have been chosen before it. We call this characteristic of their behaviour *serial dependency*. In our second study, we specifically investigate whether this characteristic is in line with what humans do by exploring more complex machine-learned models incorporating serial dependency features. We find that performance is not increased by giving the machine-learner access to information about already-included attributes and the *discriminatory power* of the referring expression under construction. This provides further support for the view that serial dependency is not a characteristic of human reference production. Consequently, algorithms incorporating this characteristic are not likely to be good models of human referring behaviour.

Alignment and conceptual pacts

Starting with the early work of Krauss and Weinheimer (1967) and Carroll (1980), a strand of research in psycholinguistics has explored how a speaker's form of reference to an entity is impacted by the way that entity has previously been referred to in the discourse or dialogue. Brennan and Clark and their colleagues (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Metzing & Brennan, 2003) found that speakers form *conceptual pacts* about the way in which they refer to objects in dialogue. They argue that conceptual pacts come about as a result of (possibly very simple) listener models maintained by the conversational participants, which are dependent on *common ground* being established between them. The implication of much of this work is that one speaker introduces an entity by means of some description, and then (perhaps after some negotiation) both conversational participants share this form of reference, or a form of reference derived from it, when they subsequently refer to that entity. Pickering and Garrod (2004) argue for a more mechanistic model whereby automatic priming accounts for the fact that a conversational participant will often adopt the same semantic, syntactic and lexical alternatives as the other party in a dialogue. This accounts for the fact that both initial and subsequent references are influenced by the previous discourse.

Although the terminology and exact mechanisms that enable alignment and conceptual pacts are debated in the field, we will refer to this work collectively as the *alignment model* of reference for the purpose of the studies presented in this paper, as the evidence for the existence of these phenomena is uncontroversial. The majority of work on alignment has focussed on the syntactic and lexcial levels of language production. For example, in Branigan, Pickering, McLean, and Cleland's (2007) experiments, participants reproduced syntactic constructions that were previously used by a different speaker describing another target referent. Lexical alignment was, for instance, found by Weiß, Pfeiffer, Schaffranietz, and Rickheit (2008), who had people play a 'jigsaw map game' involving physical objects with more than one possible name. Over time, their participants aligned on the lexical terms they used to describe the objects they were referring to repeatedly. However, we are concerned with the mechanisms that guide speakers in their choice of semantic content for a referring expression, rather than with lexical or syntactic aspects of referring expressions. In this regard, the recent work by Goudbeek and Krahmer (2012) on priming effects at the semantic level is particularly relevant. They provided evidence that speakers align the semantic content of their first reference to an object with that of previous descriptions for other objects. The participants in their study were more likely to use a dispreferred attribute to describe a target referent if this attribute had recently been used in a description by a confederate. In particular, this was the case even though the previous and current target objects did not share the same value for the dispreferred attribute. This indicates that the observed alignment effect was indeed taking place at the level of semantic attributes, such as colour, size or orientation, and not at the level of attribute values or lexical items, such as *blue*, *large* or *left-facing*.

It should be pointed out that people have been found to align their references both with previous references to the same target referent and with references to other entities. We call these two types of alignment *within-item alignment* and *cross-item alignment*. Evidence for conceptual pacts is mostly concerned with people entraining on names for the *same* object (Brennan & Clark, 1996; Metzing & Brennan, 2003). However, other work has shown that speakers can also be influenced in their referring behaviour by previous references to *other* objects. In particular, Goudbeek and Krahmer (2012) used a cross-item setting. Pickering and Garrod (2004) also argue that speakers align their references both with those produced by their conversational partner and with those they have previously uttered themselves. Buschmeier, Bergmann, and Kopp (2009) termed these two phenomena *other-alignment* and *self-alignment* in their computational implementation of the lexical aspects of Pickering and Garrod's model, which they tested on Weiß et al.'s (2008) data. Our machine-learned models

take into account both cross- and within-item alignment as well as both self- and other-alignment.

Constructive approaches to reference production

In contrast to the psycholinguistic models, work in computational linguistics and natural language generation over the last 20 years has adopted what we refer to here as a *constructive approach* to reference production. Much of this work takes as its starting point the characterisation of the problem expressed by Dale (1989). Dale considered the task of a referring expression generation algorithm to be concerned with deciding what properties of an entity should be mentioned in order to distinguish that entity from others in the current context with which it might be confused. Early work was concerned with *subsequent* reference in *discourse*, inspired by Grosz and Sidner's (1986) observations on how the attentional structure of a discourse made particular referents accessible at any given point. More recently, attention has shifted to *initial* reference. This shift has been driven in large part by the availability of purpose-built collections of one-shot *distinguishing descriptions*, such as the TUNA Corpus (van Deemter, Gatt, van der Sluis, & Power (2012), and a number of shared tasks that make use of these corpora (Gatt & Belz, 2010). Whether concerned with initial or subsequent reference, the construction of distinguishing descriptions with a focus on semantic content selection has consistently been a key consideration in this body of work.

The classic REG algorithms and their descendants follow a common schema, depicted in Figure 1. As long as the referring expression under construction does not uniquely identify the target referent, these algorithms repeat two steps: (1) add another attribute to the referring expression; (2) re-compute the set of distractors that remains. The *distractor set* is the set of entities that the target referent could be mistaken for given the referring expression constructed so far. At the beginning of the process, when the referring expression does not yet contain any attributes to distinguish the target from the distractors, the distractor set is usually taken to be the set of objects that are visually available in the target referent's environment. In a discourse context, the distractor set can also contain other objects that have recently been mentioned. This is in line with recent psycholinguistic findings that speakers use attributes that distinguish the target referent from visually close or recently mentioned potential distractors (Beun & Cremers, 1998; Brown-Schmidt & Tanenhaus, 2008). In particular, the use of relative attributes, such as size, has been shown to be dependent on the presence of distractors which differ from the target in this attribute (see, e.g., Brown-Schmidt & Konopka, 2011; Brown-Schmidt & Tanenhaus 2006; Sedivy, 2003).

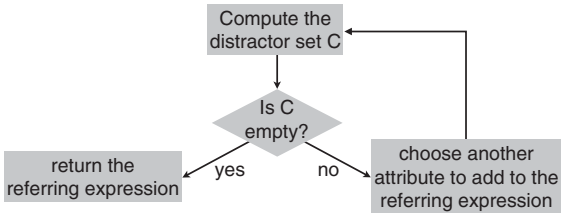


Figure 1. The traditional REG model.

Many algorithms are centred around the concept of the *discriminatory power* of an attribute or a set of attributes. Discriminatory power is determined by the number of distractor entities that can be ruled out by using an attribute or a set of attributes. More formally, it is defined as the ratio between the total number of distractors and the number of distractor entities which have a different value for a given attribute or set of attributes than the target referent. Given a set of attributes A and a distractor set C, it is computed as

$$disc(A) = \frac{|\mathrm{C} - [\![A]\!]|}{|\mathrm{C}|} \quad (1)$$

where $[\![A]\!]$ denotes the set of entities for which all attributes in A are true.

The stopping criterion in these algorithms is an empty distractor set; in other words, they are done as soon as a fully distinguishing referring expression A with $disc(A) = 1$ has been found. The main difference between the different algorithms that have been proposed lies in the attribute selection step. Dale's (1989) Greedy Algorithm attempts to find the 'best' attribute with the highest discriminatory power at that point; Dale and Reiter's (1995) Incremental Algorithm selects attributes in accordance with a pre-determined preference order over those attributes, and Krahmer, van Erik and Verleg's (2003) graph-based algorithm chooses the attribute that minimises the overall cost of the referring expression based on a pre-determined cost function over the attributes.

However, what most traditional REG algorithms have in common is that they can only include one attribute at a time, and the decision of whether an attribute is included or not hinges on the current distractor set. The Greedy Algorithm always chooses the attribute which rules out most of the remaining distractors, while the Incremental Algorithm only includes attributes that rule out at least one of the remaining distractors. Because the current distractor set is determined by the attributes that have already been included, this introduces a form of what we call *serial dependency* in the functioning of these algorithms. A

range of other reported approaches which are based on these two algorithms follow the same principles (e.g., Dale & Haddock, 1991; Gardent, 2002; Kelleher & Kruijff, 2006; Krahmer & Theune, 2002; de Lucena & Paraboni, 2008; van Deemter et al., 2012).

Another characteristic shared by all traditional REG algorithms is the fact that they are limited in the number of attributes they can include in a referring expression *redundantly*. A redundant attribute is one which could be omitted without rendering the referring expression ambiguous. So, for instance, the attribute colour in the referring expression *the large green building* is redundant, if *the large building* would have sufficed to fully distinguish the target referent from all possible distractors. A referring expression containing one or more redundant attributes is usually said to be *overspecified*. The traditional algorithms' tendency to avoid overspecification is partly due to their stopping criterion, which dictates that they cease adding attributes to a referring expression once it is fully distinguishing. A second reason lies in the fact that they only choose attributes that rule out at least some distractors. They would therefore never include an attribute shared by all objects in a given scene, nor would they choose two attributes that both rule out the same set of distractors. In the latter case, once one of those attributes has been included, the other one does not rule out any further distractors and will therefore not be subsequently considered.

It should be noted that most computational REG algorithms, including the Greedy, Incremental and graph-based algorithms, do not guarantee the generation of descriptions without redundant attributes. Under certain circumstances they do produce overspecified descriptions. In particular, the fact that the Incremental Algorithm sometimes includes redundant attributes, just as humans do, has often been touted as a sign of its potential cognitive plausibility.[1] Despite this, REG algorithms were originally rarely tested against human behaviour. A recent trend towards empirical evaluation in REG has begun to make up for this omission (Gatt & Belz, 2010; Guhe, 2012; Viethen & Dale, 2006; van Deemter et al., 2012). In these evaluations the Incremental Algorithm as well as the graph-based algorithm were able to replicate human descriptions reasonably well (Gatt & Belz, 2010; Viethen & Dale, 2006; van Deemter et al., in 2012). However, these evaluation efforts were limited to simple static scenes, rather than natural discourse, and typically only compared the performance of REG algorithms to each other; they did not test alternative psycholinguistic models of reference.

The referring expression production models we train and test in this paper are not exact equivalents of one specific constructive REG algorithm. Instead we endeavour to use machine-learning features to capture and test the *underlying assumptions* that many of these algorithms share. The main assumption of the constructive approach is that the content of a referring expression is dependent on the differences between the target referent and a set of distractor objects with which it might be confused. For existing computational algorithms, discovering these differences involves comparisons to each potential distractor, which is computationally expensive and might therefore seem cognitively implausible. However, we would argue that the idea that humans take the characteristics of surrounding objects into account at least to some extent when choosing content for referring expressions remains intuitively appealing. The advantage of the machine-learning approach we take is that it can learn heuristics such as 'if the majority of the distractors share the colour of the target, do not use it.' So, the resulting model can incorporate this assumption without having to painstakingly make each comparison, just as a human does not have to obtain an exact count in order to see if a scene is dominated by a certain colour.

The second assumption of the constructive approach that we test is that attributes are chosen in a serial fashion whereby each attribute's use is dependent on which other attributes have already been chosen. This again seems to necessitate computationally expensive mechanisms re-computing the discriminatory power of attributes in every step. However, the Incremental Algorithm in particular simplifies this mechanism by presupposing the order in which attributes are considered and only requiring a check to determine whether at least one distractor is ruled out by each attribute. Again, this is one of the reasons why it has been considered more cognitively plausible than its predecessor, the Greedy Algorithm. In testing this assumption we therefore follow the Incremental Algorithm in presupposing an order on the available attributes. Of course, we test all possible orders in order to achieve a thorough assessment of the plausibility of serial dependency.

The third assumption that the constructive approach makes is that, when a fully distinguishing description has been found, no more attributes will be added to it. This assumption presupposes a serial approach to the consideration of attributes and is implicitly tested together with the serial dependency assumption in our second study.

Models combining construction and alignment

The first researchers to look at how the constructive approach and the alignment approach might be integrated and compared to each other were Jordan and Walker (2000, 2005), who developed an early

machine-learning approach to content selection. They explored the validity of different psycholinguistic models of reference production, including a combination of Dale and Reiter's (1995) Incremental Algorithm and Grosz and Sidner's (1986) model of discourse structure, the conceptual pacts model of Clark and colleagues (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986), and the intentional influences model developed by Jordan (2000). This study found that the conceptual pact and intentional influences models were more important in shaping people's referring expressions than Grosz and Sidner's notion of discourse structure. However, the referring expressions in their data set often had functions other than identification. The corpus from which they were taken (the Coconut Corpus, Di Eugenio, Jordan, Thomason, & Moore, 2000) consists of dialogues between participants whose task was to jointly furnish a living space given limited budgets and differing furniture inventories. Therefore, in many cases certain attributes of a furniture item are mentioned in order to convince the partner that this item should be bought, or to confirm that the item has the given attribute, rather than to simply enable the listener to identify the referent, which is the main function of the referring expressions explored in this paper.

Gupta and Stent (2005) instantiated Dale and Reiter's (1995) Incremental Algorithm with a preference ordering that favoured the attributes that were used in the previous mention of the same referent. In a second variant, they even required these attributes to be included in a subsequent reference irrespective of whether they were helpful in discriminating the target from other entities at that point. They found that models that incorporated alignment and partner effects were better able to mimic the human reference behaviour in their test sets than Dale and Reiter's (1995) original Incremental Algorithm. However, departing from most other works on REG they extended the task to include ordering of the attributes in the surface form. This makes it hard to compare their model to other approaches which are not concerned with attribute ordering (including the models tested in this paper).

In a previous paper (Viethen, Zwarts, Dale, & Guhe, 2010), we presented a rule-based system that addressed a specific instance of the problem we consider more broadly here: we singled out the first references to entities by the second speaker ('second speaker initial references') and attempted to reproduce these using a system based on Dale and Reiter's (1995) Incremental Algorithm. When constructing a reference to a given entity, the system was able to take into account the reference history for that entity to a limited extent. We found that, for this particular type of referring expression, the system was not able to outperform a baseline that simply copied the previous mention of the target referent.

Guhe (2012) discusses two cognitive models, implemented in ACT-R (Anderson, 2007), that simulate the production of referring expressions in the iMAP task (the same corpus we use here). In particular, this work compares the performance of a model that is based on the Incremental Algorithm with a 'template' model that makes the decision of whether to include an attribute in a referring expression purely on the basis of the attribute's utility in the task environment. Consistent with the findings we report in this paper, the template model provides a better fit to the human data. However, it is also specifically geared towards the properties of the iMAP task environment, namely that the colour attribute is less reliable for identifying referents than another, map-specific attribute (which we will call the other attribute below). The models demonstrate that the Incremental Algorithm does not fit the data as well as the template model, because it generates uniquely distinguishing referring expressions and because it considers attributes for inclusion in the referring expression in a fixed order (what we call here *serial dependency*). Both models learn from feedback about whether the generated referring expression was successful. The model based on the Incremental Algorithm extends the original algorithm by this ability to adapt to feedback, but it overpredicts the frequency with which distinguishing expressions are made and underpredicts the frequency of overspecified referring expressions. The template model chooses attributes for referring expressions purely based on its current estimate of an attribute's utility, which is learned from feedback about successful use.

The experimental framework

The approach we follow in this paper is to investigate human referring behaviour by building computational models that perform the same referring task, and then comparing the behaviour of the models to that of human speakers. Because the different models that we test are based on the same considerations that also underlie the psycholinguistic alignment models and computational constructive approaches, we are able to compare and combine these considerations with each other in our models. Based on differences in performance in replicating the human data, we can then draw conclusions about which model's underlying assumptions play a more important role in human reference production.

The particular problem we focus on is the selection of the attributes that should be used in a referring

expression. This problem is tackled by the *Conceptualiser* module in Levelt's (1989) model of language production, and generally known as *content selection* for referring expressions in the Natural Language Generation literature. We can think of each referring expression as being a linguistic realisation of a collection of attribute–value pairs, or *properties*, which apply to the target referent. For example, the corresponding set of attribute–value pairs for the noun phrase *the purple fish* is ⟨colour: purple, type: fish⟩. The same set of properties would correspond to the slightly differently realised noun phrase *the fish that's purple*. As we will see in the following section, the values for the attributes of each object in our domain are trivial to determine, because each object has one unique value for each attribute which can be found in a database representing the domain. Therefore, we can abstract away one step further from the surface form of the referring expression in our data by ignoring the values and concentrating only on the attributes. This leads us to the concept of the *content pattern* of a referring expression: the collection of attributes that are used in that instance. For example, the content pattern of our example noun phrase *the purple fish* is ⟨colour, type⟩. The noun phrase *the greenish alien* is also associated with the same content pattern ⟨colour, type⟩, although the target referent of this expression clearly has different values for the same attributes. In the studies discussed in this paper, we investigate how well different automatically learned models can replicate the content patterns found in a large set of human-produced referring expressions.

By abstracting away from the exact linguistic realisation of each referring expression and instead concentrating on the semantic content alone, we reduce the overall complexity of the machine-learning problem, because the set of possible attribute–value pairs for each object to be described is much smaller than the set of possible fully realised noun phrases. As a consequence of this abstraction, questions regarding prenominal modifier ordering, lexical choice and syntactic realisation are beyond the scope of this paper. By taking the extra abstraction step from sets of attribute–value pairs to content patterns we further reduce the complexity of the problem. In addition, this has a further, conceptual, advantage to do with the alignment phenomena we are interested in modelling: by representing each referring expression as a content pattern, we are able to train models that capture semantic cross-item alignment, where speakers reuse attributes but not necessarily values in descriptions of different objects (as observed by Goudbeek & Krahmer, 2012). If each referring expression was represented as a set of attribute–value pairs this would not be possible, as our models could in that case only pick up on patterns of use for complete attribute–value pairs.

In Study 1, we train a number of simple models which incorporate aspects of the computational constructive approach, the alignment approach or both. In these models, the inclusion of each attribute is treated as independent of the other attributes. This independence is achieved by using what we call an *Attribute-based* method in our machine-learned models, an approach we have argued previously (Dale & Viethen, 2010). It involves training a separate *decision tree* for each attribute that occurs in the data. Once trained, each of these decision trees makes a binary choice as to whether to include the attribute it was trained for. The decisions are based on information about the situation in which the referring expression to be generated is occurring. These attribute-specific choices are then used to build a complete content pattern by including those attributes which were chosen by their model. Figure 2(a) illustrates how three decision trees, one each to decide about the inclusion of colour, pattern and type, contribute to the final content pattern. A subsequent realisation module can then combine the example output in the figure, ⟨pattern, type⟩, with information about the pattern and type values of the target referent from the domain database and realise it as, for example, *the striped fish*. This last step is, however, not part of our investigation.

In Study 2, we go on to explore more complex models, which chain the attribute-specific decision trees together in a series. In those models, the choice to include each attribute can be made dependent on information about which of the attributes that occurred earlier in the chain have already been included. The flow of this information from one decision tree to subsequent ones is indicated by dashed arrows in Figure 2(b). Analogous to Figure 2(a), the solid arrows indicate the contribution each decision tree makes to the content pattern. The example output in Figure 2(b) could then be realised as, for example, *the green triangle* or *the green triangular shape*. We use these *chained models* to investigate the validity of the serial dependency characteristic of many existing REG algorithms. How this works in detail will be explained below in the section on Study 2. In contrast to the chained models from Study 2, we call the models used in Study 1 *unchained models*.

The individual decision trees in each model are trained using the C4.5 algorithm (Quinlan, 1993), implemented as J48 in the Weka toolkit (Witten & Frank, 2005), with a pruning confidence threshold of 0.25.[2] This machine-learning algorithm takes as input a set of example instances of human-produced referring expressions, each characterised by a feature vector containing information about the situation in which

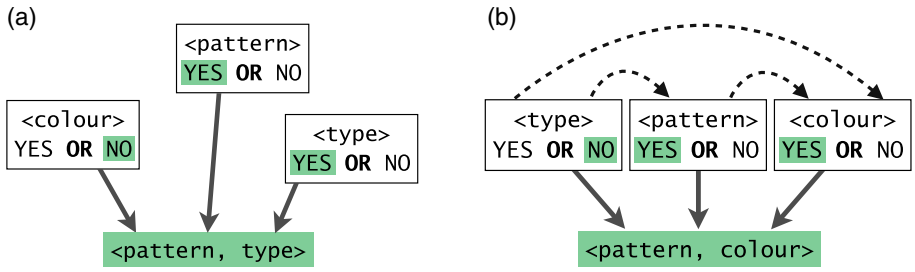


Figure 2. The structure of Attribute-Centric models.
Note: Each box is a decision tree choosing whether the indicated attribute should be used or not. The outcome of these decisions then contributes to the content pattern to be generated (solid arrows). (a) An example of an unchained model, as used in Study 1. The attribute-specific decision trees contribute to the content pattern independently from each other. (b) An example of a chained model, as used in Study 2. Each decision tree receives input information from the previous trees, indicated by the dotted arrows.

the referring expression occurred. In our case, the feature vector includes a range of information about the discourse situation, the visual stimulus presented to the speakers, the speakers themselves, and so on. This information is explained in detail in subsequent sections. Based on the set of examples, or training instances, the learning algorithm constructs a decision tree from the root up. At the beginning all training instances are at the root. Step by step, the learning algorithm splits the set of training instances according to the values that they have for one certain feature. For example, when building a tree to decide whether colour should be used it might make the first split according to a temporal feature: all instances which occurred in the first half of the dialogue are grouped in one new leaf of the decision tree, all instances from the second half in another one. Then it goes on to split the data at one of the two new leaves according to another information feature, for example whether the previous referring expression in the dialogue contained colour; and so on. Figure 3 shows an easily readable example of a decision tree that could have resulted from these first few steps. The rectangular nodes contain features according to which the tree was split at this node, and the oval leaves contain the decision that the tree returns for each instance that gets sorted into this leaf. At each step in the building phase, an information gain criterion is used to determine at which leaf a split should be made and according to which feature: the learner always introduces that split which maximises the homogeneity of the instances at each leaf in terms of whether colour was used for them by the human speakers. In other words, it attempts to produce a tree that sorts as few example instances as possible into leaves at which the final decision as to whether to use colour or not is different from the ones the human speakers actually made.

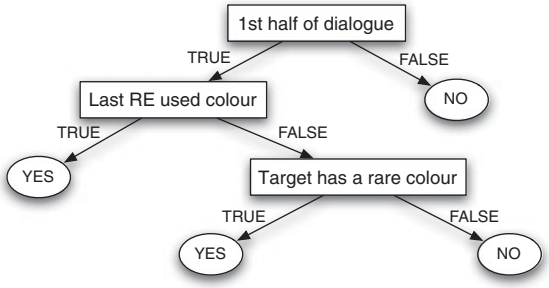


Figure 3. A tree that uses three information features to decide whether colour should be used in a referring expression.
Note: This decision tree is only an example for explanatory purposes and is not based on actual data.

There are different ways to determine the importance of a feature or a set of features in a decision tree. If only a small selection of all features is chosen for a decision tree by the machine-learning algorithm, it is clear that these chosen features are of high importance. However, the number of features included in a tree can be very large and features can be used multiple times in different branches of the tree, which makes it more difficult to assess their importance. (The largest tree in our experiments contains 2199 nodes.) It is theoretically possible to re-compute the overall information gain that each individual feature contributes to a tree; however, the most straightforward method to find out how important a certain feature or set of features is for the behaviour of a given tree is to perform a *feature ablation study*, whereby the feature or features of interest are removed from the overall feature set available to the machine learner. A new tree is trained without access to these features, and its ability to correctly predict the outcome for a number of test instances is compared to that of the original tree including the features. If removing the features of interest leads to a significant decrease in prediction accuracy, we know that they play an important role in the behaviour of the original tree. Because access to these features allowed the original tree to better

match the behaviour of the human participants who produced the training and test data, this outcome also points to a high likelihood that the factors captured in these features played an important role in the behaviour of these human participants. Reversely, if removal of a set of features does not result in a decrease in performance, we know that they had no impact on the behaviour of the original model and are also not likely to play a large role in the behaviour of the human participants underlying that model. Similar to Jordan and Walker (2005), who took advantage of this feature ablation technique to assess the validity of their intentional influences model, we make use of it in order to assess the importance of factors based on the assumptions underlying the Constructive and the Alignment approaches. We do not report on the importance of individual features. An exhaustive ablation study for all features would require too much space while distracting from the main focus of this paper.

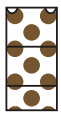
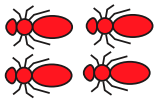
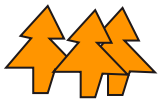
Referring expressions in the iMAP Corpus

Our explorations are based on the iMAP Corpus (Guhe, 2012; Louwerse et al., 2007), a collection of 256 two-person dialogues. It was collected and annotated by Max Louwerse, Ellen Gurman Bard, Mark Steedman and colleagues as part of the now concluded iMAP project.[3]

Thirty-two participant pairs contributed to the iMAP Corpus in an experiment akin to the original Map Task (Anderson et al., 1991). The participants were all undergraduate students at the University of Memphis. The two members of a participant pair each saw a map of the same environment; one participant in each pair was the *instruction giver*, the other was the *instruction follower*. The instruction giver's map showed a path winding between a large number of landmarks; this path was not visible on the instruction follower's map. The task was for the instruction giver to describe the path to the instruction follower in such a way that the instruction follower could draw it onto his map.

The maps were constructed from a set of eight types of landmarks, which are distinguishable by their colour and one other attribute. The eight types are grouped into four pairs according to their second distinguishing attribute, which we call the other attribute. For each of the four other attributes there is one animate and one inanimate landmark type. Table 1 shows the two landmark types for each other attribute and one example for each type. The four other attributes are pattern, shape, number and kind. The landmarks that are distinguishable by pattern are of type fish and car; they can be plain, dotted, checkered, or horizontally or vertically striped. The shape landmarks are of type alien or traffic sign, and they can be round, triangular, rectangular or hexagonal. Bugs and trees come in little clusters of one to five, where each cluster counts as one

Table 1. The other distinguishing attributes and the landmark types which use them.

	Animate		Inanimate	
Attribute	Type	Example	Type	Example
Pattern	Fish	 Checkered	Car	 Dotted
Shape	Alien	 Triangular	Traffic Sign	 Round
Number	Bugs	 Four	Trees	 Three
Kind	Bird	 Ostrich	Building	 Church

landmark. This makes number the other distinguishing attribute for bugs and trees. Finally, distinguishable by kind are bird and building landmarks. Birds have one of the kinds eagle, penguin, robin, ostrich, swan or owl; and buildings have one of the kinds house, church, castle, steeple, shop and apartment block. Before the start of each dialogue, participants were prompted with the types of the landmarks they were about to encounter on the next set of maps, in order to reduce the variability in the otherwise unrestricted dialogue task. In addition to the direct attributes of the landmarks, participants used spatial relations to describe them. These relations describe the location of the target referent relative to another landmark, to the path or to parts of the map.

To make the route description task more demanding, and to ensure that the participants had to engage in a dialogue, there are always a few discrepancies between the instruction giver's and the instruction follower's maps. (The map task has been used frequently to elicit clarification requests.) Some landmarks differ in the value of their type, colour or other attribute, and some landmarks might be missing. Furthermore, the instruction follower's map has some ink damage, obscuring the colour of some landmarks. The ink damage is either 'orderly' (one large ink blot) or 'disorderly' (several smaller blots). Each map uses one of the eight landmark types as the main type. Half of the maps show only landmarks of the main type (single condition); the other half shows also landmarks of other types, although the main landmark type is still predominant (mixed condition). Even in the mixed condition, all landmarks appearing on the same map are either animate or inanimate. Figure 4(a) shows an example pair of bird maps; in this case, they are mixed with disorderly ink blots on the instruction follower's map. Figure 4(b) shows the instruction giver's single fish map.

Each participant pair contributed eight dialogues, each covering one map. They switched roles after four maps, so that each participant played the role of instruction giver in four dialogues and that of instruction follower in the other four. This resulted in 256 dialogues. Figure 5 shows the start of one of the dialogues based on the maps in Figure 4(a). The dialogues were hand-annotated for referring expressions to the intended landmarks (marked in bold in Figure 5), and each referring expression was coded with the attributes it contains. In this way, each referring expression is associated with its content pattern. For example, for a referring expression such as *the blue penguin* we have the annotation ⟨colour, kind⟩. The actual values for the attributes found in a content pattern for a particular landmark can be obtained from a database containing information about all landmarks on a given map. This makes it trivial to recover the full semantic content (e.g., ⟨colour:blue, kind:penguin⟩) from a given content pattern by looking up the value for each attribute contained in it.

For our studies, we removed from the data any annotation that was not concerned with the landmark's type and colour, a spatial relation or the landmark's other distinguishing attribute. For example, 'semantically empty' head nouns (*rectangular thing* or *that set of*

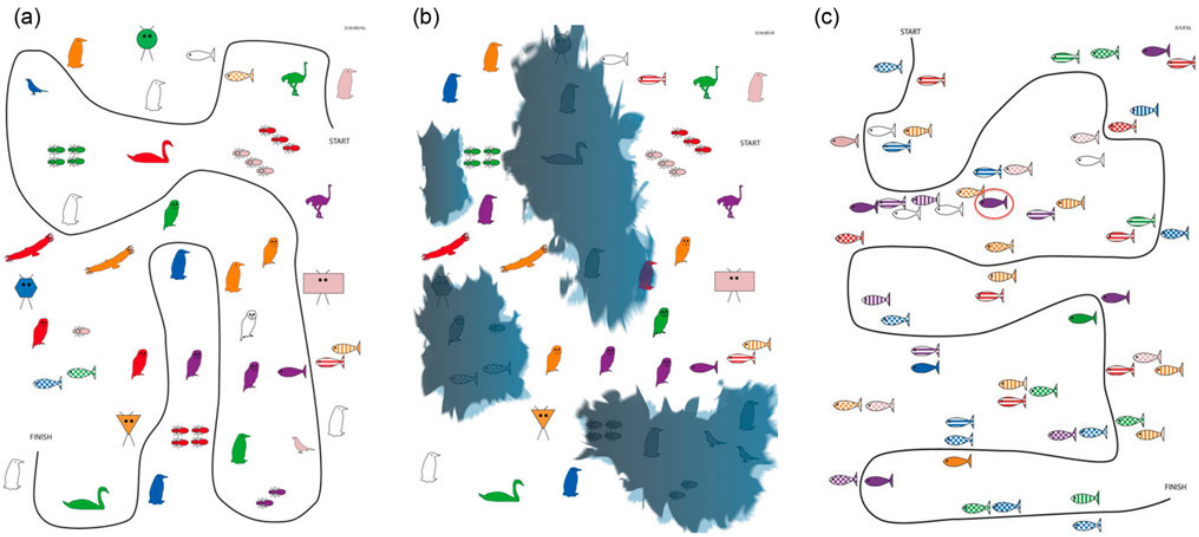


Figure 4. Three example maps. (a) An example pair of maps: the mixed bird maps with disorderly ink damage on the instruction follower's map. (b) The instruction giver's single fish map.

1 IG: do you see start
2 IF: on the right side
3 IG: do you see an **pink penguin** and a **green ostrich right above that**
4 IF: mhmm
5 IG: alright you're gonna go between **those** and uh just go like right past **them** and stop
6 IF: go to the left
7 IG: go right between the uh **pink penguin** and the **green ostrich** right between you're gonna be close to the **pink penguin**
8 IF: okay
9 IG: that make sense
10 IF: yeah
11 IG: and then stop
12 IG: you're gonna do one of your curvy corner things and go up like you're gonna be about a quarter of an inch above the **ostrich**'s head
13 IF: okay
14 IG: uh do you see a **solid fish** and then a **dotted fish like kind of below and to the right of it**
15 IF: of the **penguin** i mean of the
16 IG: no to the left of the **ostrich**
17 IF: the **ostrich** i mean
18 IG: yeah
19 IF: um i have i have i have **one** i have **two fish** but **they**'re different than
20 IG: what are **they**
21 IF: i have **white one** and **red striped one**
22 IG: okay where's the **red striped one** - like pretty much even with **it** or right below **it**
23 IF: with the **ostrich**
24 IG: with the **white fish**
25 IF: um the **white fish** is above them all
26 IG: okay - do you have a **circular alien**
...

Figure 5. The start of a dialogue about the maps in Figure 4(a).
Note: The referring expressions annotated in the corpus are marked in bold face.

three fish), demonstrative determiners incorrectly tagged as pronouns (*that blue alien*), as well as the mention of size (*a big purple rectangle*) were filtered out. Ordinal numbers that were annotated as the use of the number attribute were re-tagged as spatial relations, as these usually described the position of the target within a line of landmarks (*the third fish*, for the lowest (orange) fish in the vertical group of three fish towards the bottom of Figure 4b).

As a result of the removal of these annotations, 2785 referring expressions had no annotation left; we removed these instances from the final data set. We also do not attempt to replicate the remaining 5552 plural referring expressions or the 3062 pronouns found in the corpus.[4]

Furthermore, for the studies presented here, we also excluded 2586 descriptions that made use of a spatial relation. It is not possible to determine the discriminatory power of a spatial relation, because there is no definite way to establish the number of distractor objects that could be described using the same spatial relations as a given target referent. However, as mentioned above, the discriminatory power of each attribute is a crucial ingredient of the constructive approach. The data used for our previous experiment (Viethen et al., 2011a) nonetheless included relational descriptions, which might have disadvantaged the constructive approach in the experiments in that paper. Excluding relational descriptions from the data set, as we do here, is the more rigorous solution for this problem.

The most fine-grained level of annotation we have available only indicates whether a spatial relation was used with respect to another landmark (e.g., *below the ostrich*), the route (e.g., *to your left*) or parts of the map (e.g., *in the middle of the screen*). To be able to determine the discriminatory power of such a spatial relation, we would need to have a formal annotation of the actual value of the relation, information about the relatum (e.g., attributes of the related landmark, or the part of the map) and a database that holds information about how many distractor objects could also be described by this exact spatial relation. These resources are neither available at this point nor feasible to construct. As this is the kind of information that is crucial for the constructive approach, expecting our models to deal with relations without this information would put the constructive approach at a disadvantage. Furthermore, without information about the discriminatory power of all attributes it would be impossible to apply the method we use in Study 2 to test the serial dependency assumption. We aimed to use the same

data set for both studies in order to be able to directly compare the results, and therefore omitted the relational descriptions also in Study 1.

This leaves 20,141 referring expressions, a very large set compared to the collections of referring expressions used in related research, such as Jordan and Walker's (2005) data set of 393 instances or the set of 1765 referring expressions used by Gupta and Stent (2005). Our final data set comprises 5936 initial references and 14,205 subsequent references. Table 2 lists the seven content patterns that occur in the final data set in order of frequency.

The basic features

The number of factors that might impact on the content of a referring expression in a dialogic setting is very large. Attempting to incorporate all these factors into parameters for a rule-based model, and then experimenting with different settings for these parameters, would be prohibitively complex. Instead, we capture a wide range of factors as features that can be used by a machine-learning algorithm to train models associated with different feature subsets. In particular, we are here interested in comparing and combining feature sets based on the constructive tradition of referring expression generation to feature sets reflecting the considerations at the heart of alignment-based approaches to reference stemming from psycholinguistic research.

The features we extracted from the data set are listed in Tables 3–5.[5] They fall into three main subsets: *Constructive* features capture factors that the constructive approaches to computational referring expression generation take account of; *Alignment* features capture factors that we would expect to play a role in the psycholinguistic models of alignment and conceptual pacts; and theory-*External* features capture factors that are not based on either of these two approaches.

The Constructive features (Table 3) can be further separated into *Visual* and *Discourse* features. Visual features capture aspects of the visual appearance of the objects around the target referent, while Discourse features capture the properties of objects that have recently been mentioned in the dialogue. These two categories of features are based on research which has shown that speakers in comparable tasks restrict the referential domain to objects that are located close to the focus of attention or have been recently mentioned (Beun & Cremers, 1998; Brown-Schmidt & Tanenhaus, 2008).

For the Visual Constructive features, we operationalise the spatial proximity criterion by defining a circle around each landmark; every other landmark whose centre point falls within the circle is considered a *visual distractor*. We tested a number of different ways of setting the size of the visual context set and came to the conclusion that there is no straightforward method to determine exactly how many distractors should be taken into account (Viethen, Dale, & Guhe, 2011c). We therefore had to rely on an informal visual analysis of the maps similar to Guhe (2012) to set the size of the circles defining the visual referential domain. For each map type a different radius is defined in such a way that on average each landmark on the maps of this type has six distractors. For most landmarks this means that at least all other landmarks in 'direct line of sight' are included in the visual context. This resulted in slightly larger circles on the maps with fewer objects, and smaller circles on the more dense maps. We also introduce features that are only concerned with the landmark that is closest to the target referent (Closest_Same_*Att* and Dist_Closest), as it can be assumed

Table 2. The seven content patterns that occur in the data set used in Study 2, ordered by frequency.

Content pattern	Count	Proportion (%)
⟨other⟩	7561	37.5
⟨other, type⟩	5975	29.7
⟨other, colour⟩	2364	11.7
⟨other, colour, type⟩	1954	9.7
⟨colour⟩	1029	5.1
⟨type⟩	662	3.3
⟨colour, type⟩	596	3.0
Total	20,141	

Table 3. The Constructive feature set.

Constructive features (Visual)	
Count_Vis_Distractors	The number of visual distractors
Prop_Vis_Same_*Att*	The proportion of visual distractors with the same value for *Att* as the target
Dist_Closest	The distance to the closest visual distractor
Closest_Same_*Att*	Has the closest distractor the same value for *Att*?
Dist_Closest_Same_*Att*	Distance to the closest distractor with the same value for *Att* as the target
Cl_Same_type_Same_*Att*	Has the closest distractor of the same type also the same value for *Att*?
Constructive features (Discourse)	
Count_Disc_Distractors	The number of other landmarks mentioned since the last mention of the target
Prop_Disc_Same_*Att*	The proportion of recently mentioned landmarks for which *Att* was used and which have the same *Att* as the target

that no matter what the actual size of the visual context is, the closest neighbouring object will be part of it. These features are similar to the approach to dealing with the visual context set in the take on the Incremental Algorithm described by Guhe (2012), which only considers one distractor at a time from the distractors (1) within a circle of a fixed radius around the landmark and (2) in direct line of sight.

For the Constructive Discourse features, the discourse proximity criterion is operationalised by taking into account only objects that have been referred to since the last mention of the current target referent, or since the beginning of the dialogue, if the current mention is an initial reference. These objects are considered the *discourse distractors*. In determining the set of discourse distractors, we also took into account referring expressions that were excluded from the final set of descriptions to be replicated (see previous section).

Of particular interest are the features Prop_Vis_Same_*Att* and Prop_Disc_Same_*Att*. They record the proportion of other landmarks, within the set of visual or discourse distractors, which have the same value for the attribute *Att* as the current target referent. This proportion approximates how useful the attribute is in distinguishing the target from the distractors overall, without requiring an exact count of distractors sharing the given attribute with the target. This can also be considered an approximation of the visual salience of the attribute, which is usually taken to be determined by how unique an attribute value is in the visual context. In Study 2 we introduce 'chaining' versions of these features, which are used to chain attribute-specific decision trees into series in models incorporating the serial dependency characteristic of some constructive REG algorithms. These chaining versions are updated every time an attribute is included into the referring expression under construction, which means that they directly represent the discriminatory power of the attribute under consideration. Having unchained and chained versions of these attributes allows us to compare the different impacts that visual salience and discriminatory power have on the content of referring expressions.

The Alignment feature set (Table 4) can be divided into the *Recency* and *Frequency* features. Recency features keep track of how long ago a given attribute was last used for the same target referent or for any object in the same dialogue. The assumption is that the more recently a speaker has heard an attribute in a priming reference, the more likely he is to reuse it. For example, Dist_Last_Mention_Utts records how many utterances ago the target referent was last mentioned, and Last_Mention_*Att* records whether the attribute *Att* was used in this mention. These two features are concerned with within-item alignment, and they capture both self-alignment and other-alignment, since they do not distinguish whether it was the same speaker who uttered the last mention or not. If the identity of the speaker of the last mention should be an important factor, then the machine learner can draw in the Last_Men_Speaker_Same feature. Frequency features count how often an attribute has been used, based on the assumption that more frequent previous use leads to more robust priming. These features only take into account the context up until the point of the current reference to be produced, as only previous references can have a priming effect. This is different from the frequency-based preference orders commonly used in implementations of the Incremental Algorithm, which take counts over the complete training set and thereby capture more general preferences rather than priming.

Table 4. The Alignment feature set.

Alignment features (Recency)	
Last_Men_Speaker_Same	Who made the last mention of the target?
Last_Mention_*Att* was *Att*	Used in the last mention of the target?
Dist_Last_Mention_Utts	The distance to the last mention of the target in utterances
Dist_Last_Mention_Res	The distance to the last mention of the target in referring expressions
Dist_Last_*Att*_LM_Utts	The distance in utterances to the last use of *Att* for the target
Dist_Last_*Att*_LM_Res	The distance in referring expressions to the last use of *Att* for the target
Dist_Last_*Att*_Dial_Utts	The distance in utterances to the last use of *Att*
Dist_Last_*Att*_Dial_Res	The distance in referring expressions to the last use of *Att*
Dist_Last_RE_Utts	The distance to the last referring expression in utterances
Last_RE_*Att*	Was *Att* mentioned in the last referring expression?
Alignment features (Frequency)	
Count_*Att*_Dial	The number of times *Att* has been used in the dialogue so far
Count_*Att*_LM	The number of times *Att* has been used for the target so far
Quartile	The quartile of the dialogue in which the referring expression was uttered
Dial_No	The number of dialogues already completed +1
Mention_No	The number of previous mentions of target +1

We include features that are only concerned with within-item alignment (Count_*Att*_LM) as well as features capturing both within- and cross-item alignment (Count_*Att*_Dial). Note that because of our focus on the semantic level of reference production, the alignment features record the use of attributes rather than attribute values. However, the landmark-specific features (those with names containing _LM_) necessarily confound these two levels because a given landmark does not change its value for a given attribute between mentions. These features are likely to also capture some lexical and even syntactic alignment, as it is probable that the more often a given attribute of a landmark is used, the more often it is used in a certain surface form.

Again, the values of all features take into account referring expressions that were excluded from the final data set. For example, if three referring expressions to other objects occurred since the last mention of the current target referent, the value of Dist_Last_Mention_REs is 4, even if the intervening references did not contain any of the attributes we are trying to model and are therefore not part of the final set of references on which we train and test our models. If a feature is not applicable in a given instance, as is for example the case with Dist_Last_Mention features for the first mention of a landmark, a special value was recorded. This special value is the number 1000 for numerical features and the string 'N/A' for categorical features. This makes it possible for the machine learner to pick up on reference strategies specific to these cases.

Also grouped under Frequency features are considerations about how long the participants have already been involved in the experiment, because, for example, a low frequency of use at the start of a dialogue might have less impact than a low frequency of use at the end of a dialogue. As Guhe and Bard (2008) show, the frequency with which attributes are used changes over the course of a dialogue as well as over the course of the entire experiment.

The External features (Table 5) are composed of the *Map* features, the *Speaker* features and the *LMprop* features. Map features capture design characteristics of the map the current dialogue is about; Speaker features capture the identity and role of the current speaker; and LMprop features capture the inherent visual properties of the target referent.

Table 5. The External feature set.

Map features	
Main_Map_type	The most frequent type of landmark on this map
Main_Map_other	The other attribute of the most frequent type of landmark
Mixedness	Are other landmark types present on this map?
Ink_Orderliness	Shape of the ink blot(s) on the instruction follower's map
Speaker features	
Dyad_ID	Identifier of the participant pair
Speaker_ID	Identifier of the person who uttered this referring expression
Speaker_Role	Was the speaker the instruction giver or the instruction follower?
LMprop features	
other_Att	The other attribute of the target
*Att*_Value	The value for each *Att* of the target
*Att*_Difference	Was the value for the target's *Att* different between the two maps?
Missing	Was the target missing one of the maps?
Inked_Out	Was the target inked out on the instruction giver's map?

Evaluation set-up

We used a 70–30% split to divide the data into a training set and a test set. In addition to the main prediction class content pattern, the training–test set split was stratified for Speaker_ID and Quartile to ensure that training and test set contained the same proportion of descriptions from each speaker and each quartile of the dialogue. We report the results for initial and subsequent reference separately because it seems likely that the two approaches impact differently on the different data subsets. In particular, within-item alignment can by definition only apply for subsequent references, while people might be more likely to carefully construct a distinguishing description for an initial reference. Table 6 shows the exact training–test set splits.

We use Accuracy and average Dice score for our comparisons; these are the most commonly used measures in the REG literature (see, for example, Gatt, Belz, & Kow, 2008). Accuracy reflects the proportion of cases for which a model produces output identical to that found in the test set. The Dice coefficient (Dice, 1945) is a set-comparison metric that delivers values ranging from 0 to 1. A Dice score of 0 signifies that two sets have no common members and 1 signifies that the

Table 6. The sizes of training and test sets for the three data subsets.

	Initial references	Subsequent references	All references
Training set	4140	9909	14,038
Test set	1796	4296	6103
Total	5936	14,205	20,141

sets are identical. Given two sets of attributes, A and B, their Dice similarity is computed as

$$\text{Dice}(A,B)=\frac{2\times|A\cap B|}{|A|+|B|}. \quad (2)$$

This gives some measure of the overlap between two referring expressions, assigning a partial score if the two sets share attributes but are not identical. The Accuracy of a full content selection system is the proportion of test instances for which it achieves a Dice score of 1, signifying a perfect match between the predicted content pattern and the content pattern found in the human test data. The difference between the two metrics can be demonstrated by a simple example. Suppose a human speaker has described an object with a referring expression containing the content pattern ⟨colour, other, type⟩, and one of our models produces the content pattern ⟨colour, type⟩ for the same instance. In this case, the Accuracy of the model for this instance is 0, because it did not perfectly replicate the content pattern of the human speaker. However, the Dice score would be (2 * 2/5 =) 0.8, because the content pattern produced by the model overlaps considerably with that corresponding to the human reference.

Study 1: comparing unchained models

This feature ablation study examines to what extent people carefully construct referring expressions (the constructive approach) and to what extent they align with previous references (the alignment approach), or whether perhaps both processes are involved in reference production. In this study we investigate simple models of the constructive perspective not incorporating the serial dependency characteristic of some algorithms. We use our training set to build different reference production models based on the different feature subsets described above. We then test these models on the held-out test set of referring expressions, in order to see which subset of features best captures the referring behaviour of the speakers in our corpus.

The models

We trained a number of different models using the decision-tree learning algorithm based on different subsets of the features described above. As already mentioned, we used an Attribute-based approach whereby a separate decision tree is trained to decide, for each of the three attributes type, colour and other, whether it should be used or not. The decisions of these three trees are then combined to form a full content pattern. Depending on the features included in a model, it follows either the findings in the psycholinguistic literature or the assumptions of existing algorithmic approaches, or it combines the two with each other and with theory-external considerations.

AllFeatures uses decision trees trained on all features, thereby drawing on theory-external considerations as well as those of the alignment and constructive approaches.

Constructive uses decision trees trained on the Constructive features alone, assuming that only the considerations of constructive approach play a role in reference production.

Alignment uses decision trees trained on the Alignment features alone, assuming that only the considerations of the alignment approach play a role in reference production.

External uses decision trees trained on the External features alone, assuming that neither constructive nor alignment considerations play a role in reference production and that instead other, theory-external, factors alone shape referring expressions in dialogue.

Alignment+External uses decision trees trained on all but the Constructive features, assuming that the considerations underlying the alignment approach combined with theory-external factors can fully account for human reference behaviour.

Constructive+External uses decision trees trained on all but the Alignment features, assuming that the considerations underlying the constructive approach combined with theory-external factors can fully account for human reference behaviour, and

Constructive+Alignment uses decision trees trained on all but the External features, assuming that both alignment and constructive considerations play a role, while theory-external factors do not.

In addition, we test three simple baseline models, which did not require training as such, but nonetheless take some aspects of the training data into account:

HeadNounOnly generates only the property that is most frequently realised as the head noun for the target, which is kind for birds and buildings and type for all other landmarks. This is a form of 'reduced reference' strategy.

MajorityClass generates the content pattern most commonly used in the training set.

RepeatLast represents a very simplistic alignment approach, which is applicable only for subsequent reference. It generates the same content pattern that was used in the previous mention of the target referent.

Results

Table 7 compares the performance of the three baselines and the models based on the seven feature subsets. The Majority baseline predicts the content pattern ⟨other⟩ on all three data sets. In our analyses we use χ^2

Table 7. Performance of the models and baselines tested in Study 1.

	Initial references		Subsequent references		All references	
	Acc	Dice	Acc	Dice	Acc	Dice
HeadOnly	19.5	59.7	26.1	52.6	23.9	54.4
Majority	28.7	70.9	41.4	70.0	37.4	69.8
RepeatLast	–	–	41.6	56.8	–	–
Constructive	55.4	85.4	55.0	80.8	54.0	81.5
Alignment	62.8	87.8	60.6	82.9	62.4	84.7
External	67.8	90.2	61.1	84.3	61.8	85.3
Alignment+External	**72.2**	**91.7**	65.2	85.5	**68.7**	**87.9**
Constructive+External	69.2	90.5	61.9	84.5	63.1	86.1
Constructive+Alignment	67.3	89.5	60.8	83.4	63.4	85.4
AllFeatures	**72.3**	**91.6**	**66.0**	**86.2**	**68.2**	**88.1**

Note: Accuracy values are given in%. In each column, the performance of the best models is marked in boldface.

to compare Accuracy scores and the Wilcoxon Signed-Rank test to compare Dice scores with a Bonferroni adjusted $\alpha = 0.001$.

The table indicates that the learned systems outperform all three baselines. The statistical analyses confirm that the performance of all models is significantly better than that of the highest performing baseline on each data set both measured by Accuracy ($\chi^2 > 153.8$, df = 1, $p << 0.001$) and by Dice ($Z > 6.9$, $p << 0.001$).

A comparison of the Alignment model and the Constructive model shows that the content patterns resulting from the Alignment model are a better match to the human-produced patterns both in terms of Accuracy ($\chi^2 > 20.0$, df = 1, $p << 0.001$ for all three data sets) and Dice on the subsequent and the complete data sets ($Z > 3.3$, $p < 0.001$), but not the initial referring expressions ($Z = 1.5$, $p = 0.13$). Interestingly, even the External model outperforms the Constructive model (Accuracy: $\chi^2 > 32$, df = 1, $p << 0.001$, Dice: $Z > 5.2$, $p << 0.001$, for all three data sets).

A comparison of the AllFeatures model to each the Constructive+External and the Alignment+External models again shows an advantage for the Alignment features over the Constructive features: removing the Constructive features from the complete feature set (resulting in the Alignment+External model) has no significant impact compared to the best-performing AllFeatures model on any of the data sets (Accuracy: $\chi^2 < 0.5$, df = 1, $p > 0.46$, Dice: $Z < 1.9$, $p > 0.47$), while removing the Alignment features from the complete feature set (resulting in the Constructive + External model) significantly hurts performance compared to AllFeatures on the subsequent and full data set in terms of Accuracy ($\chi^2 > 15.7$, df = 1, $p << 0.001$) and on the full data set in terms of Dice ($Z = 4.9$, $p << 0.001$).

The good performance of the Constructive+External model on the initial references seems to make sense intuitively because one might assume that the first time a speaker refers to an object he is more likely to carefully construct a reference, while alignment might become more important in subsequent references. However, the numbers suggest that this good performance is due to the External features, which by themselves already do just as well on this data set as the combined Constructive+External model. In fact, on the initial references, the performance of the External model is also not significantly lower than that of the best-performing AllFeatures model (Accuracy: $\chi^2 = 8.7$, df = 1, $p > 0.003$, Dice: $Z = 2.8$, $p > 0.005$).

Discussion

In this study, we have captured both the psycholinguistic alignment approach and the traditional computational view of REG via sets of features for machine learning. Additionally, we defined a number of theory-external features. We then trained decision trees on different feature subsets and compared their ability to model a held-out test set. Using this approach, we have one main finding about the influence that the considerations underlying these different views have on the content of human referring expression production.

The results suggest that considerations at the heart of constructive REG approaches do not play as important a role as those postulated by alignment-based models for the selection of semantic content for subsequent referring expressions. We have demonstrated that a model using all these features to predict content patterns in references in a dialogue setting delivers an Accuracy of 68.2% and a Dice score of 88.1. However, we found that the features based on

constructive REG considerations do not contribute as much to this score as those based on the alignment approach, and that removing the Constructive features alone does not significantly hurt the performance of a model based on alignment and theory-external features. Interestingly, this is not only the case for subsequent reference, but also for initial reference, where one might expect that distinguishing from the visual context would be of more importance. We will discuss this finding in more detail in the General Discussion. The Accuracy and Dice scores were, however, higher for all models on the initial references than on the subsequent references, indicating that initial references are easier to predict based on our set of features.

We note that the Accuracy scores achieved by our learned systems are similar to the best numbers previously reported in the REG literature and are much higher than chance level for this task.[6] Even in the arguably much simpler non-dialogic domains of the REG competitions concerned with pure content selection for initial references, the best-performing system achieved only 53% Accuracy (see Gatt et al., 2008). The most comparable approach, the rule-based system we recently applied to a subset of the data used here, was not able to outperform a RepeatLast baseline at 40.2% Accuracy and an average Dice score of 67.0 (Viethen et al., 2010).

Study 2: assessing serial dependency

The results of Study 1 suggest that the considerations underlying traditional REG algorithms do not have a large impact on the content of subsequent referring expressions in map task dialogues. In this second study, we investigate the psychological validity of a further characteristic of many traditional REG approaches that was not modelled in Study 1: as discussed in the introduction, many algorithms follow a principle that we call *serial dependency*, whereby attributes are selected for inclusion one at a time, and the decision to include each attribute is dependent on the discriminatory power of the set of attributes that have already been selected and of the attribute under consideration.

To investigate whether serial dependency is a property of human referring expression generation, we train *chained models*, which chain the attribute-specific decision trees into a series by using a number of *Chaining features*, and compare their performance to the simpler unchained models from Study 1. The Chaining features provide information about the discriminatory power of both the referring expression constructed so far and the attribute currently under consideration. They capture the information required for training models with the serial dependency characteristic inherent in most traditional constructive algorithms for REG which include attributes one after the other and make their decisions dependent on how distinguishing an attribute is based on the already chosen attributes. We therefore hypothesise that, if serial dependency plays a role in the generation of referring expressions, then models that use Chaining features should achieve a closer match to the human-produced data in our test set than models that only use non-chaining features.

The models

Again, we trained attribute-specific decision trees that each make a binary decision as to whether or not one of the three attributes type, colour and other should be used. The output of the three trees is then combined into a content pattern, which can be compared to the content pattern of the corresponding human-produced description in our corpus. To derive chained models, we trained 'chains' of attribute-specific decision trees for each of the six possible sequences of the three attributes. In a chained model, each decision tree is given access to Chaining features which provide information about the referring expression constructed so far and about the decisions made about the use of the previous attributes in the chain. Of central importance for these features is the concept of discriminatory power defined above. The three types of Chaining features that we use are as follows:

1. DP_*Att* represents the discriminatory power of attribute *Att* at the time at which *Att* is considered for inclusion. We use two variants of this feature, one pertaining to visual distractors (the surrounding landmarks on the map) and one to discourse distractors (landmarks that have recently been mentioned in the dialogue). Most existing REG algorithms, including the Incremental Algorithm and the Greedy Algorithm, make their decision as to whether to include an attribute *Att* dependent on its current discriminatory power.

The DP_*Att* features are related to the Constructive features Prop_Vis_Same_*Att* and Prop_Disc_Same_*Att*, but with the important difference that their values are computed at run time, taking into account that the size of the total distractor set might already be reduced by attributes that have already been included. This makes the DP_*Att* features chaining versions of the two non-chaining Constructive features.

2. DP_RE records the discriminatory power of the referring expression built so far. Existing algorithms stop adding attributes as soon as the discriminatory power of the referring expression reaches 1, which means that all distractors are ruled out. Again, we use both visual and discourse variants of this feature.

3. Incl_*Att* features record whether attribute *Att* has been included in the referring expression for all

attributes that precede the current one in the chain. This feature captures an aspect of serial dependency that is not usually represented in traditional REG approaches: it allows the machine learner to pick up on patterns of one attribute only being used if another one is also mentioned.

We built seven pure chained models based on combinations of the Chaining features, and one which included all Chaining features and all non-chaining features:

1: uses decision trees trained on the DP_*Att* features only.
2: uses decision trees trained on the DP_RE features only.
3: uses decision trees trained on the Incl_*Att* features only.
1+2: uses decision trees trained on the DP_*Att* and the DP_RE features.
1+3: uses decision trees trained on the DP_*Att* features and the Incl_*Att* features.
2+3: uses decision trees trained on the DP_RE features and the Incl_*Att* features.
1+2+3: uses decision trees trained on all Chaining features.

We trained a further seven chained models by adding all three Chaining features to the different non-chained models from Study 1.

Considering the similarity of the DP_*Att* features to the non-chaining Prop_Vis_Same_*Att* and Prop_Disc_Same_*Att* features, we also compared these two types of features directly. To this end we trained one further model, which can be compared directly to the chained model **1**:

1NC: uses decision trees trained on the Prop_Vis_Same_*Att* and Prop_Disc_Same_*Att* features.

Results

Table 8 shows the Accuracy and Dice scores achieved by the 14 chained models in replicating the human-produced data. Again, we use χ^2 to compare Accuracy scores and the Wilcoxon Signed-Rank test to compare Dice scores, both with a Bonferroni adjusted $\alpha = 0.001$.

We tried all possible orders in which the three attributes colour, other and type can be chained, but report only the result of the best-performing order for each model on each data set. The table shows that by themselves the Chaining features perform relatively badly compared to the mixed models also including non-chaining features. The difference between the best pure chained model (1+2+3) and the worst mixed model (1+2+3+Constructive) is statistically significant both in terms of Accuracy ($\chi^2 > 62.5$, df = 1, $p << 0.001$) and Dice ($Z > 6.7$, $p << 0.001$) on all three data subsets.

Comparing these results to those from Study 1 in Table 7 shows, first, that the models based only on Chaining features also perform worse than all unchained models. The best chained model (1+2+3) performs significantly worse than Constructive, the worst unchained model (Accuracy: $\chi^2 > 55.8$, df = 1, $p << 0.001$, Dice: $Z > 6.6$, $p << 0.001$, on all three data subsets). Secondly, we see that adding Chaining features to unchained models does not increase the performance in terms of Accuracy ($\chi^2 < 5.0$, df = 1, $p > 0.02$, for all pairs of unchained models and their chained counterparts, on all three data subsets) and with three the exceptions in terms of Dice ($Z < 2.7$,

Table 8. Accuracy (in%) and Dice score achieved by the chained models tested on the three data subsets in Study 2.

	Initial references		Subsequent references		All references	
	Acc	Dice	Acc	Dice	Acc	Dice
1	39.9	78.7	43.7	73.0	41.1	73.1
2	42.0	78.7	41.8	70.4	38.5	72.8
3	39.0	78.8	41.4	70.0	37.4	69.8
1+2	42.0	78.7	44.3	74.0	41.5	73.7
2+3	42.0	78.7	41.8	70.4	38.5	72.8
1+3	39.9	78.7	44.6	73.5	41.3	73.6
1+2+3	42.9	79.8	44.3	74.0	41.4	73.8
1+2+3+Constructive	56.1	85.5	55.6	80.8	54.7	81.8
1+2+3+Alignment	65.6	88.9	62.0	83.5	63.5	84.9
1+2+3+External	68.3	90.2	62.8	85.1	63.8	86.3
1+2+3+Alignment+External	**72.3**	**91.8**	**66.4**	**86.5**	**68.8**	**88.0**
1+2+3+Constructive+External	69.3	90.4	63.1	85.2	64.1	86.4
1+2+3+Constructive+Alignment	67.9	89.8	62.0	83.6	64.0	85.6
1+2+3+AllF	**72.5**	**91.6**	**66.4**	**86.0**	**68.6**	**87.9**

$p > 0.007$ for all models on all data sets with the exception of the Alignment model on the initial data, $Z = 3.3$, $p < 0.001$, the Alignment+External model on the subsequent data, $Z = 3.9$, $p << 0.001$, and the External model on the full data, $Z = 3.7$, $p << 0.001$). In particular, the best-performing unchained models, Alignment+External and AllFeatures, achieve 68.7 and 68.2% Accuracy, respectively, on the set of all references; the chained versions of these models, 1+2+3+Alignment+External and 1+2+3+AllFeatures, achieve an almost indistinguishable 68.8% and 68.6%.

Table 9 compares the performance of Model 1, based on the chaining features DP_*Att*, to that of Model 1NC, based on the equivalent non-chaining features. The features used for these two models are almost identical, with the one crucial difference that Model 1 computes the discriminatory power of an attribute only at the point at which it decides whether to include it, while 1NC determines the attributes' discriminatory power independent of the referring expression under construction. This comparison shows that using the non-chaining version of this feature outperforms the chaining version (Accuracy: $\chi^2 > 21.5$, df = 1, $p << .001$, Dice: $Z > 3.6$, $p << 0.001$, on all three data subsets).

Discussion

The main results of this study (Table 8) demonstrate that the Chaining features do not contribute significantly to an accurate model of human production of referring expressions: all non-chaining models from Study 1 outperform all models based only on Chaining features; and combining chaining and non-chaining features in one model does not increase performance. This lends support to the view that the characteristic of serial dependency that is central to the many REG algorithms, such as the Incremental Algorithm and its decendants, does not accurately reflect the way in which humans generate referring expressions. It appears that factors other than discriminatory power, such as alignment with the previous discourse, better explain the referring behaviour of human speakers in task-oriented dialogue.

Table 9. Comparison of the Accuracy (in %) of chained and unchained features representing the discriminatory power of the three attributes in Study 2.

	Initial references		Subsequent references		All references	
	Acc	Dice	Acc	Dice	Acc	Dice
1	39.9	78.7	43.7	73.0	41.1	73.1
1NC	47.6	81.4	51.8	78.0	49.2	78.3

The non-chaining features Prop_Vis_Same_*Att* and Prop_Disc_Same_*Att* represent the discriminatory power of the individual properties of the target referent independent of the referring expression under construction. This is similar to the notion of an attribute's visual salience, which is usually taken to be determined by how fast an object's value for this attribute is being made available by the human perceptual system and how well this value differentiates the object from the surrounding objects. The results from Table 9 therefore indicate that visual salience might be of more importance in the choice of attributes for subsequent reference than dynamically computed discriminatory power. In essence, people do not seem to compute the discriminatory power of an attribute when they decide whether to use it in a repeated reference, but rather make a decision based on the attribute's overall visual salience within the visual context of the intended referent.

These results provide further support for our earlier proposal that the attributes in a referring expression might be chosen independently and in parallel, based on simple scene analysis and alignment rather than on a more computationally expensive selection processes involving serial dependency and discriminatory power (Dale & Viethen, 2010). They are also consistent with the findings of Guhe (2012), which show that an attribute's overall utility in the task environment influences how frequently it is chosen.

Error analysis

An important question to ask is how wrong the models really are when they do not succeed in perfectly matching a human-produced reference in our test set. It might be that they choose a completely different set of attributes from those included by the human speaker; however, the Accuracy score also counts as incorrect any set that partly overlaps with the reference found in the test set. The Dice score gives us a partial answer to this question, as it assigns a score that is based on the size of the overlap between the attribute set chosen by the model and that included by the human speaker. The fact that all our models achieved Dice scores much higher than their Accuracy scores shows that they only rarely get it completely wrong.

A more fine-grained analysis can be obtained by examining for each model how many of the content patterns it produced contained a subset of the human reference for the same instance, how many were a superset of the attributes included in the human reference, how many had another form of partial intersection with the human reference, and how many had no commonality with the human reference. Examples for these different types of intersection are

	model output		human reference	
	content pattern	example realisation	content pattern	example realisation
Subset:	⟨other⟩	*the rectangle*	⟨other, type⟩	*the rectangular alien*
Superset:	⟨colour, other⟩	*the purple church*	⟨other⟩	*the church*
Other intersect:	⟨other, type⟩	*the round traffic sign*	⟨colour, type⟩	*the green traffic sign*
No overlap:	⟨other⟩	*the robin*	⟨colour, type⟩	*the blue bird*

Figure 6. Examples for the four different ways in which the models' output can intersect with content patterns found in the human references. Note: The surface form realisations in these examples are extrapolations.

given in Figure 6. Table 10 shows the proportions of subsets, supersets, other intersections and non-overlapping content patterns produced by each model on the full data set. All models make most of their mistakes by producing a subset or a superset of attributes in the human reference. Other forms of intersection and content patterns that have no common attributes with the human-produced referring expression in the corpus are much rarer. This is in line with the relatively high Dice scores our models achieved.

Most notably, the badly performing models trained only on Constructive and Chaining features often undergenerate with respect to the corresponding human references, meaning that the models based on the constructive approach often chose too few attributes. In particular, the models based purely on Chaining features produce a very large percentage of content patterns that are subsets of those contained in the referring expressions the human participants used in the same situations. It is likely that this brevity in the expressions produced by the Constructive and Chaining models is due to the assumptions underlying them that also lead traditional REG algorithm to aim for brevity: taking into account the discriminatory power of each attribute, considering attributes in a serial fashion and taking into account the discriminatory power of the already assembled set of attributes. The fact that the better-performing alignment-based models do not make the mistake of undergeneration as frequently as the constructive models suggests that our Alignment features capture aspects of why people decide to use these additional attributes. The relatively high rate of overgeneration for the models containing Alignment features shows that this tendency to add additional attributes sometimes goes too far.

The two models containing Constructive but no Alignment or External features (Constructive and 1+2+3+Constructive) also have a higher incidence of

Table 10. The proportion (in %) of test instances for which each model in Study 1 produced a subset or a superset of the human reference or had some other form of intersection or no overlap with it.

	Accuracy	Subset	Superset	Other intersect	No overlap
1	41.1	**40.6**	7.3	2.2	8.8
2	38.5	**33.3**	14.3	5.3	8.6
3	37.4	**51.1**	0.0	0.0	11.5
1+2	41.5	**40.7**	7.3	2.4	8.1
2+3	38.5	**33.3**	14.3	5.3	8.6
1+3	41.3	**41.0**	7.3	2.2	8.3
1+2+3	41.4	**40.6**	7.6	2.5	8.0
1+2+3+Constructive	54.7	**18.9**	16.4	5.2	4.8
1+2+3+Alignment	63.5	13.2	**15.8**	2.9	4.6
1+2+3+External	63.8	11.9	**19.1**	2.2	3.1
1+2+3+Alignment+External	68.8	10.2	**16.2**	1.9	2.9
1+2+3+Constructive+External	64.1	12.6	**18.2**	2.1	3.0
1+2+3+Constructive+Alignment	64.0	13.3	**16.2**	2.5	4.0
1+2+3+AllFeatures	68.6	10.7	**15.7**	2.1	3.0
Constructive	54.0	**20.1**	16.0	4.8	5.1
Alignment	62.4	13.5	**17.0**	2.4	4.7
External	61.8	13.2	**19.2**	2.2	3.6
Alignment+External	68.7	10.6	**15.7**	1.9	3.1
Constructive+External	63.1	13.8	**17.9**	2.2	3.1
Constructive+Alignment	63.4	13.5	**16.7**	2.3	4.1
AllFeatures	68.2	10.5	**16.8**	1.8	2.6

Note: Only the numbers for the full data set are shown. The most common error for each model is marked in boldface.

other-intersect and no-overlap errors than any model containing Alignment or External features. However, these types of errors are together less than half as common as subset errors, and the Constructive models' rate of superset errors is very similar to that of the best-performing Alignment+External and AllFeatures models. This leaves undergeneration as the main culprit for the comparatively bad performance of the Constructive models, with other-intersect and no-overlap errors contributing to a smaller extent.

A second observation from Table 10 supports our conclusion from both studies that the features based on the considerations underlying traditional REG approaches (Constructive and Chaining features) are not as influential in the reference production process as the Alignment and also the External features: We have already seen that the models based only on Constructive and Chaining features produce a larger number of subsets than supersets of the human references. For the models only using Alignment and External features, on the other hand, this ratio is reversed. The table shows that all other models, which contain a mix of those groups of feature sets (e.g., 1+2+3+Alignment or Constructive + External), produce a ratio similar to the Alignment and External models: they are also more likely to produce a superset of the human reference than a subset. This indicates that the behaviour of these 'mixed' models is dominated by the Alignment and External features, while the Constructive and Chaining features do not contribute substantially to their behaviour.

General discussion

In this paper we have used a machine-learning approach in order to evaluate and compare different models for the selection of semantic content for reference production. We used machine-learning features that were drawn from two main models stemming from different research fields. The first is what we have termed the *constructive approach*, which is embodied in many referring expression generation algorithms developed within computational Natural Language Generation. The second is what we call the *alignment approach*, which originates from research in psycholinguistics into the effects that priming and conceptual pacts have on the way people refer. By couching the considerations underlying these two different approaches in terms of machine-learning features we were able to train models taking into account the considerations underlying either one of the two approaches or both at the same time.

In our first study we found that, while a model using all features at once performed equally well, removing only the features associated with the constructive approach did not hurt the performance of the model. Furthermore, a model using only the constructive features was outperformed by all other learned models. The implication of this is that people seem to generate referring expressions with little regard for the visual and discourse context captured in these features, or at least that the influence of these features is masked by other factors (such as alignment) that play a bigger role. So, we conclude that the view that people produce referring expressions by deliberately constructing distinguishing descriptions should at least be considered suspect. This seems to be a plausible position if we look only at subsequent reference, as we did in Viethen et al. (2011a): once an entity has been introduced into the discourse, perhaps how it is referred to subsequently depends more on the preceding discourse than it does on the visual context at the time of reference. Indeed, once an entity has been referred to, the description that has been constructed 'factors in' the visual context, and so any subsequent reference to that entity does not require re-computation of the description; referring to the entity in the way that it was referred to before should still do the job (unless, of course, the context has changed in some relevant way). Such a model has the twin appeals of being both more computationally efficient, and consistent with explanations based on the alignment approach.

However, perhaps surprisingly, our current results show that the constructive approach also seems to play no role in content selection for *initial* reference, where one might expect that distinguishing from the other objects in the context would be important, as no previous reference can be reused. So what is going on here? Intuition suggests that, in real-world scenes, we do take account of the distinguishing ability of our referring expressions; when we describe an intended referent, we do not do so blindly without considering whether the referring expression might be confusing or ambiguous. But our data suggest that, at least in the dialogue scenarios we have looked at, this is not the case. Even the initial references in the iMAP Corpus can be modelled more accurately by a system following the alignment approach than by one built on the assumptions of the constructive approach.

The straightforward conclusion from these results is that people indeed do not carefully construct initial referring expressions, but rely on priming and alignment mechanisms to choose their semantic content. This is in line with the evidence for semantic cross-item alignment presented by Goudbeek and Krahmer (2012). They found that people are likely to reuse attributes they have recently heard even when they are describing a new object and even if the value for the

attribute they are reusing is different from the one in the prime. Related to this is recent evidence for general adaptation to factors of the task environment over time. Guhe (2012) found that people's use of the colour attribute in the iMAP Corpus was more strongly influenced by their learning over time that colour is not always reliable (captured by our Alignment and External features) than by the discriminatory power of colour in a given local environment (captured by our Constructive features). We have to take into account that the speakers also have a rather demanding task to solve, which will require most of their attention and might not leave much cognitive resources for the relatively routine task of referring. If speakers do not have enough 'processing power' left for the careful construction of referring expressions, they are likely to fall back on the computationally less demanding Alignment model.

Some redemption for the constructive approach might be found in the simplicity of the domain underlying the iMAP data: one might argue that this does not reflect the complexity of typical real-world visual scenes, and that the complex mechanisms we think are required for reference production more generally are simply not necessary in these simple scenarios. Rather than compute a reference that takes account of the context, the participants in the iMAP task perhaps (unconsciously) recognised that the scenes are simple enough to use referring expressions that are not carefully computed on the basis of context, even when referring to a landmark for the first time. In a more complex environment, we would then expect that people will either devote more cognitive resources to the construction of referring expressions in order to maintain the same success rate, or suffer an increase in specific breakdowns of referential success. In that case, participants might take longer at completing their task, or their success rate at completing the task might decrease. This raises a methodological issue with work aimed at verifying models of human reference production in both psycholinguistics and Natural Language Generation: perhaps it is a flawed assumption that we can test models in relatively simple but arguably also quite abstract and unnatural domains (such as the ones researchers often use in order to control the number of factors playing a role) and hope that the results will generalise to more complex and realistic scenarios.

From a technical point of view, it has to be noted that our machine-learned models are able to make use of missing alignment information as well: they can pick up on patterns in the speakers' reference behaviour that specifically apply to situations in which little or no information is available about previous referring expressions that could be aligned to. The most extreme example of such a situation is the very first reference uttered in a dialogue. For this reference, none of the alignment features provide any information about previous references. Instead these features have a special value indicating that the feature is not applicable (see the section on *The Basic Features* above). The machine learner can therefore learn rules that specifically deal with these cases. In a way, no information is also information from the perspective of the machine learner. For example, the decision tree making the choice whether colour should be used in the Alignment+External model for initial references includes the rule in Example (3). It indicates that for instances in which the Last_RE_COL feature is not applicable (i.e., the very first reference in a dialogue where no previous referring expression exists), colour should not be used.[7] Furthermore, the Alignment+External model can make use of the theory-external features for cases where no alignment information is available. For example, in addition to many alignment features, the Alignment+External model for initial references makes use of the Role, Speaker_ID, Main_Map_Att, as well as the Mixedness features, in order to determine whether type should be used. This means that the rules these models use for initial reference, and in particular for the very first reference in a dialogue, reflect people's base strategies for content selection in the absence of precedents to align to.

$$\text{Last\_RE\_COL} = \text{N/A} : \text{no.} \tag{3}$$

Our second study delivers a further blow to possible claims of cognitive plausibility for the computational constructive approaches. Here we found that the characteristic of serial dependency inherent in some of the most classic REG algorithms does not seem to play a role in the way humans construct referring expressions. The serial dependency characteristic implies that attributes for a target referent are chosen one after the other based on how well they distinguish the target from the other objects in the context, taking account of the other attributes which have already been chosen. The fact that this does not seem to be what people do is in line with the results from the first study, which also found that people pay much less attention to features of the context than is assumed by the computational constructive approaches.

Our error analysis showed that the poor performance of the models based on features representing the computational REG approaches was mainly caused by them choosing subsets of the attributes that the human references contained. This points to the fact that human speakers tend to overspecify their descriptions more than the models did. The fact that classic REG algorithms, such as the Incremental Algorithm, are not able to include as much redundant information as humans do has been lamented in the literature before

(Koolen, Goudbeek, & Krahmer, 2011; Viethen & Dale, 2006; van Deemter et al., 2012), and there have been calls for computational models that pay less attention to the attributes' discriminatory power and instead use simple, computationally cheap, heuristics.

In particular, Dale and Viethen (2010) argued for an attribute-centric model, where the attributes are considered in parallel based on their visual salience as well as their discriminatory power. The unchained Attribute-based models we have used in the studies in the present paper are examples of such attribute-centric models, because here a separate decision is made for each attribute independent of the other attributes, and in principle these decisions can be made in parallel. However, our empirical results show that, at least in the dialogic setting we have examined in this paper, discriminatory power needs to take a backseat and that alignment factors are more important for such models. The significance of visual salience of attributes was confirmed by the second result from Study 2. This demonstrates that the overall prevalence of the target's value for an attribute has more impact on the likelihood of that attribute being chosen for a referring expression than the number of remaining distractors that can be ruled out by this attribute after other attributes have already been included.

A similar argument for the importance of visual salience in models of reference production was made by Koolen et al. (2011). They found that the more variation there was in a scene in terms of a given attribute, the more likely people were to include it redundantly. They argue that this is due to people following a heuristic, whereby, given a lot of variation, it is easier to simply include an attribute instead of carefully scanning the whole scene and computing its actual usefulness. Their results also showed that people sometimes use colour, an attribute particularly easy to perceive, even when it had no discriminatory power at all. A reference production model making use of quick heuristics based on visual salience is also in line with Pechmann's (1989) early eyetracking results, which demonstrated that people often begin uttering visually salient attributes, such as colour or brightness, before they have fixated on any objects other than the target referent. Our results suggest that a reference production model would not only have to include heuristics that are guided by visual salience, but that it would also have to contain heuristics for following the lead of previous references that a speaker can simply align with.

Future directions

As we have seen, even the best-performing Alignment + External model achieves far from perfect Accuracy, although the high Dice score (see Table 7) and the low rate of no-overlap mistakes (see Table 10) indicate that it rarely misses the mark completely. A number of extensions of this work are conceivable, which might help improve the performance of the models. First, it would be possible to include value-level features in the alignment models in order to fully capture priming of attribute values. At the moment only within-object priming is captured at this level. Considering that the alignment model already outperforms the constructive model, adding additional linguistic levels to the alignment model can be expected to further strengthen this effect.

A second, considerably more labour-intensive extension would be to include features to allow learning the use of spatial relations between the objects on the maps. As mentioned in the section describing our data set, relatively complex assessments as to which relationships hold between which objects would be necessary to allow the constructive approach to take relations into account properly. However, as we have argued that the Constructive features do not seem to contribute to the performance of the best models, it would be worthwhile to see how well a model combining alignment and external features would be able to capture the use of spatial relations in the full iMAP data.

Third, machine-learning features based on further annotation of the corpus for discourse-level phenomena and semantic information might lead to an overall better match between our models and the human-produced data. In particular, information about the success of previous references and about explicit grounding utterances would make it possible for the machine-learned models to take into account whether the reference to be generated is a repair for an unsuccessful earlier one or whether an earlier reference can safely be repeated or shortened. More specific for the iMAP data would be semantic information that captures whether the participants have explicitly discussed the unavailability of colour on parts of the IF's map, which could be a cue for the use of colour information to be dropped from consideration.

Fourth, it would be worth investigating whether the constructive approach would perform better if information about path and linguistic context were taken into account for the definition of the context set. At least the instruction giver, who knows the direction of the path, is likely to be influenced by this, which might lead to more oval-shaped context set circles being more appropriate. Similarly, the linguistic context of a referring expression can restrict the distractor set. For example, the distractor set for a landmark described after the sentence *Keep going until you hit a . . .* should probably only contain objects between the current position and the position of the target landmark.

Features based on these kinds of information would also require extensive additional annotation.

Conclusion

Using a machine-learning approach, we have tested a number of computational instantiations of reference production models on a large set of referring expressions taken from a corpus of dialogues in a map task domain. Two existing approaches served as inspiration for the features that we used to train our models: the psycholinguistic alignment approach and the computational constructive approach. We found that in this scenario people do not seem to construct references with the express goal of distinguishing the target referent from possible distractor items, but rather they copy content from previous references in the same dialogue when building a new referring expression. Furthermore, our results invalidate the serial dependency characteristic inherent in many traditional referring expression generation algorithms as a factor impacting on human referring behaviour, and demonstrate that visual salience has a larger effect on the content of referring expressions than discriminatory power. Based on our results we argue for an attribute-centric model of reference production in which quick heuristics based on visual salience and alignment take a paramount role over the strategic computation of content patterns based on discriminatory power and serial dependency.

Acknowledgements

Preliminary versions of the studies in this paper were presented at the Empirical Methods in Natural Language Processing conference in 2011 (Viethen, Dale, & Guhe, 2011a) and at the CogSci Workshop 'Bridging the Gap between Computational, Empirical and Theoretical Approaches to Reference' in 2011 (Viethen, Dale, & Guhe, 2011b).

Jette Viethen thanks the Netherlands Organisation for Scientific Research (NWO) for support from the VICI grant 'Bridging the Gap between Computational Linguistics and Psycholinguistics: The Case of Referring Expressions' (277-70-007). We all thank the three anonymous reviewers for their constructive comments, as well as Emiel Krahmer, Martijn Goudbeek and Albert Gatt for providing helpful feedback on an earlier version of this manuscript.

Notes

1. It is perhaps noteworthy that the article introducing this algorithm (Dale & Reiter, 1995) was in fact published in the journal *Cognitive Science* rather than a computational publication.
2. Decision trees are pruned after construction to avoid over-fitting to the training data. This usually increases their performance on unseen test data. The confidence threshold determines how severely the tree will be pruned; 0.25 is the default value provided by Weka.
3. http://www.madresearchlab.org/Media/Pages/iMAP.html
4. The additional issues that arise in generating plural references and deciding when to use pronouns considerably complicate the problem; see Gatt, 2007.
5. In these tables, *Att* is an abbreviatory variable that is instantiated once for each of the three attributes type, colour and the other distinguishing attribute of the landmark.
6. A model choosing one of the seven possible content patterns at random would only achieve an Accuracy of 14.3% (1 in 7).
7. Of course, it is hard to imagine a real-life situation with no previous context to align to at all. In particular, the first reference in an iMAP dialogue might be influenced by the previous dialogue or by the prompts given to the participants in the training sessions. However, this information was not available for our Alignment features.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe? Vol. 3 of Oxford series on Cognitive Models and Architectures*. New York, NY: Oxford University Press

Anderson, A. H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., ... Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, *34*, 351–366. doi:10.1177/002383099103400404

Beun, R.-J., & Cremers, A. (1998). Object reference in a shared domain of conversation. *Pragmatics and Cognition*, *6*(1/2), 121–152. doi:10.1075/pc.6.1-2.08beu

Branigan, H. P., Pickering, M. J., McLean, J. F., & Cleland, A. A. (2007). Syntactic alignment and participant role in dialogue. *Cognition*, *104*(2), 163–197. doi:10.1016/j.cognition.2006.05.006

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1482–1493. doi:10.1037/0278-7393.22.6.1482

Brown-Schmidt, S., & Konopka, A. E. (2011). Experimental approaches to referential domains and the on-line processing of referring expressions in unscripted conversation. *Information*, *2*, 302–326. doi:10.3390/info2020302

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*, 592–609. doi:10.1016/j.jml.2005.12.008

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science*, *32*, 643–684. doi:10.1080/03640210802066816

Buschmeier, H., Bergmann, K., & Kopp, S. (2009). An alignment-capable microplanner for natural language generation. In *Proceedings of the 12th European Workshop on Natural Language Generation* (pp. 82–89). Athens, Greece.

Carroll, J. M. (1980). Naming and describing in social communication. *Language and Speech*, *23*, 309–322.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. doi:10.1016/0010-0277(86)90010-7

Dale, R. (1989). *Cooking up referring expressions*. In Proceedings of the 27th annual meeting of the Association for Computational Linguistics (pp. 68–75). Vancouver BC, Canada.

Dale, R., & Haddock, N. (1991). *Generating referring expressions involving relations*. In Proceedings of the 5th conference of the European chapter of the Association for Computational Linguistics (pp. 161–166). Berlin, Germany.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*, 233–263. doi:10.1207/s15516709cog1902_3

Dale, R., & Viethen, J. (2010). Attribute-centric referring expression generation. In E. Krahmer & M. Theune *Empirical methods in Natural Language Generation* (Vol. 5980, pp. 163–179). Lecture Notes in Computer Science. Berlin: Springer.

van Deemter, K., Gatt, A., van der Sluis, I., & Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, *36*(5), 799–836. doi:10.1111/j.1551-6709.2011.01205.x

Di Eugenio, B., Jordan, P. W., Thomason, R. H., & Moore, J. D. (2000). The agreement process: An empirical investigation of human–human computer-mediated collaborative dialogues. *International Journal of Human-Computer Studies*, *53*, 1017–1076. doi:10.1006/ijhc.2000.0428

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*, 297–302. doi:10.2307/1932409

Gardent, C. (2002). *Generating minimal definite descriptions*. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 96–103). Philadelphia PA, USA.

Gatt, A. (2007). *Generating coherent reference to multiple entities* (Doctoral dissertation). University of Aberdeen, UK.

Gatt, A., & Belz, A. (2010). Introducing shared task evaluation to NLG: the TUNA shared task evaluation challenges. In E. Krahmer & M. Theune (Eds.), *Empirical methods in Natural Language Generation* (Vol. 5980, pp. 264–295). Lecture Notes in Computer Science. Berlin: Springer.

Gatt, A., Belz, A., & Kow, E. (2008). *The TUNA Challenge 2008: Overview and evaluation results*. In Proceedings of the 5th International Conference on Natural Language Generation (pp. 198–206). Salt Fork, OH, USA.

Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering and referential overspecification. *Topics in Cognitive Science*, *4*(2), 269–289. doi:10.1111/j.1756-8765.2012.01186.x

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

Guhe, M. (2012). Utility-based generation of referring expressions. *Topics in Cognitive Science*, *4*, 306–329. doi:10.1111/j.1756-8765.2012.01185.x

Guhe, M., & Bard, E. G. (2008). *Adapting referring expressions to the task environment*. In Proceedings of the 30th annual conference of the Cognitive Science Society (pp. 2404–2409). Austin, TX.

Gupta, S., & Stent, A. (2005). *Automatic evaluation of referring expression generation using corpora*. In Proceedings of the Workshop on using corpora for Natural Language Generation (pp. 1–6). Brighton, UK.

Jordan, P. W. (2000). *Intentional influences on object redescriptions in dialogue: Evidence from an empirical study* (Doctoral dissertation). University of Pittsburgh, Pittsburgh, PA, USA.

Jordan, P. W., & Walker, M. (2000). *Learning attribute selections for non-pronominal expressions*. In Proceedings of the 38th annual meeting of the Association for Computational Linguistics. Hong Kong, China.

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

Kelleher, J., & Kruijff, G.-J. (2006). *Incremental generation of spatial referring expressions in situated dialog*. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 1041–1048). Sydney, Australia.

Koolen, R., Goudbeek, M., & Krahmer, E. (2011). *Effects of scene variation on referential overspecification*. In Proceedings of the 33rd annual meeting of the Cognitive Science Society (pp. 1025–1030). Boston, MA, USA.

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In K. van Deemter & R. Kibble (Eds.), *Information sharing: Reference and presupposition in language generation and interpretation* (pp. 223–264). Stanford, CA: CSLI.

Krahmer, E., & van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Computational Linguistics*, *38*(1), 173–218. doi:10.1162/COLI_a_00088

Krahmer, E., van Erik, S., & Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, *29*(1), 53–72. doi:10.1162/089120103321337430

Krauss, R. M., & Weinheimer, S. (1967). Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behavior*, *6*, 359–363. doi:10.1016/S0022-5371(67)80125-7

Levelt, W. M. J. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Louwerse, M. M., Benesh, N., Hoque, M. E., Jeuniaux, P., Lewis, G., Wu, J., & Zirnstein, M. (2007). *Multimodal communication in face-to-face computer-mediated conversations*. In Proceedings of the 29th annual conference of the Cognitive Science Society (pp. 1235–1240).

de Lucena, D. J., & Paraboni, I. (2008). *Combining frequent and discriminating attributes in the generation of definite descriptions*. In Advances in Artificial Intelligence – IBERAMIA 2008 (pp. 252–261). Lecture Notes in Computer Science. Berlin: Springer.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*, 201–213. doi:10.1016/S0749-596X(03)00028-7

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110. doi:10.1515/ling.1989.27.1.89

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–226. doi:10.1017/S0140525X04000056

Quinlan, J. R. (1993). C4.5: Programs for machine learning. San Francisco, CA: Morgan Kaufmann.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23. doi:10.1023/A:1021928914454

Viethen, J., & Dale, R. (2006). *Algorithms for generating referring expressions: do they do what people do*? In Proceedings of the 4th International Conference on Natural Language Generation (pp. 63–70). Sydney, Australia.

Viethen, J., Dale, R., & Guhe, M. (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. Edinburgh, UK.

Viethen, J., Dale, R., & Guhe, M. (2011b). *Serial dependency: Is it a characteristic of human referring expression generation*? In Proceedings of the 2011 Workshop on Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference. Boston, MA, USA.

Viethen, J., Dale, R., & Guhe, M. (2011c). *The impact of visual context on the content of referring expressions*. In Proceedings of the 13th

European Workshop on Natural Language Generation. Nancy, France.

Viethen, J., Zwarts, S., Dale, R., & Guhe, M. (2010). *Dialogue reference in a visual domain*. In Proceedings of the 7th International Conference on Language Resources and Evaluation. Valetta, Malta.

Weiß, P., Pfeiffer, T., Schaffranietz, G., & Rickheit, G. (2008). Coordination in dialog: Alignment of object naming in the jigsaw map game. In H. D. Zimmer, C. Frings, A. Mecklinger, B. Opitz, M. Pospeschill, & D. Wentura, *Proceedings of the 8th annual meeting of the Cognitive Science Society of Germany* (pp. 1–17). Saarbrücken, Germany.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann.