Justifying Corpus-Based Choices in Referring Expression Generation

Helmut Horacek
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, GERMANY
helmut.horacek@dfki.de

Abstract

Most empirically-based approaches to NL generation elaborate on co-occurrences and frequencies observed over a corpus, which are then accommodated by learning algorithms. This method fails to capture generalities in generation subtasks, such as generating referring expressions, so that results obtained for some corpus cannot be transferred with confidence to similar environments or even to other domains. In order to obtain a more general basis for choices in referring expression generation, we formulate situational and task-specific properties, and we test to what degree they hold in a specific corpus. As a novelty, we incorporate features of the role of the underlying task, object identification, into these property specifications; these features are inherently domain-independent. Our method has the potential to enable the development of a repertoire of regularities that express generalities and differences across situations and domains, which supports the development of generic algorithms and also leads to a better understanding of underlying dependencies.

1 Introduction

Choices in NL generation, as geared by examples taken from a corpus, are essentially driven by observed frequencies of partial surface expressions and their co-occurrences in this corpus. Generating referring expressions (GRE) aiming at the identification of an entity or a set of entities in a situational context is the subtask addressed by most approaches in this fashion: corpora are created for the purpose of analyzing human preferences, and several GRE challenges have been conducted over some corpora and are still under way (e.g., (Gatt and Belz 2008)). By and large, this strategy leads to quite good results, the best systems performing very accurately. However, this approach has an essential drawback: it fails to capture regularities that underly the choices observed, so that they can be expressed in a somehow general form that abstracts from details of the domain and ideosyncracies of the corpus. Abstractions of this kind are a prerequisite to transfer the results obtained in the context of a corpus to similar environments or even to other domains with reasonable confidence, which is an essential goal of empirically-based approaches.

In this paper, we attempt to find out relations between task-relevant situational properties and components of the referring expressions that subjects produced for a given corpus. We formulate situational and task-specific properties, and we test to what degree they hold in a specific corpus. As a novelty, we incorporate features of the role of the task, object identification, into these property specifications; these features are inherently domain-independent. We are convinced that the resulting regularities capture facets of principled preferences in a mildly abstracted form so that they allow a reasonable transfer to other domains. Ultimately, this techniques is intended to provide an improved basis for choices in GRE.

This paper is organized as follows. We first discuss previous work, then we motivate our approach. In the main sections, we describe the ingredients in building hypothesized regularities, and we define this method in formal terms. Then we give some preliminary results. Finally, we discuss our achievements and possible impacts, and we sketch extensions and future developments.

Proceedings of Recent Advances in Natural Language Processing, pages 311–318,
Hissar, Bulgaria, 7-13 September 2013.

2 Previous Work

The task of generating referring expressions is a subtask in the traditional NL generation pipeline, the most intensively addressed one in the past decade (see (Krahmer, van Deemter 2012) for a recent overview). For a long time, there was a debate about algorithmic solutions that adequately combine computational issues with human preferences in the selection of attributes. Earlier work was characterized by featuring computational issues, such as full brevity versus the greedy heuristic (Dale, 1989), which models task properties in the search process in terms of the discriminatory power of attributes. These approaches were challenged by psychological insights, such as the role of salience (e.g., color can be perceived much quicker than other properties) and the use of redundant attributes (Pechmann 1989), a crucial issue in the GRE task. Ultimately, the dabate has been settled in favor of the incremental algorithm (Dale and Reiter 1995), which is intended to reflect these insights. The algorithms proposed have been compared in terms of their searching techniques (Bohnet and Dale 2005). The incremental algorithm contains a parameter for expressing domain-specific preferences among attributes – its instantiation has significant impact on results and quality of the expressions generated. However, the motivated specification of preferences and the attitude towards the use of redundant attributes still remain open questions.

In order to address this issue, corpora are built to examine human preferences in detail. These corpora must be the product of controled experiments, since precise evidence is needed about the situational context in which a corpus has been created. A prominent example is the TUNA corpus (Gatt, v. d. Sluis, and Deemter 2007, van Deemter et al. 2012). It comprises referring expressions from two domains: the identification of a piece of furniture resp. a person out of a set of such items, presented in a small grid. An even bigger corpus is (Guhe and Bard 2008), and two corpora based on more realistic situational 3D scenes underlying the experiments are GRE3D3 (Viethen and Dale 2008), and the bigger follow-up corpus GRE3D7 (Viethen and Dale 2011)[1] .

[1] The corpus is available for download online at www.clt.mq.edu.au/research/projects/gre3d7.

These corpora served then for the investigation of more data-oriented approaches to GRE so that they could be evaluated (Gupta and Stent 2005). Some of these corpora have been used or they are built through challenges, such as (Koller et al. 2010); competing systems try to approximate unseen examples on the basis of a corpus subset. A challenge based on the TUNA corpus is the shared task of GRE (Gatt and Belz 2008, Gatt and Belz 2010). Most participants used an adaptation of the *Incremental Algorithm*, the domain-specific parameter being modified by corpus frequencies that are accommodated by some learning algorithm. Further elements to drive choices are hard-coded rules (Kelleher and McNamee 2008), and personalized preferences (Bohnet 2008) – trials contain labels to identify the subject who produced the expression. The best systems, which used suitable learning algorithms, performed very well (see the summaries in (Belz and Gatt 2007, Gatt, Belz, and Kow 2008, Gatt, Belz, and Kow 2009)).

Apart from these challenges, a number of approaches have tried to find principles or generalities on the basis of observed data. Jordan and Walker (2005) have encapsulated the ingredients of choices in GRE in terms of rules, Viethen et al. (2010) and Viethen, Dale and Guhe (2011a) have examined the role of visual context. Finally, Viethen, Dale and Guhe (2011b) have attempted to characterize the behavior of humans in GRE: they found that the view of accommodating previous references is generally more appropriate than a purely constructive view, which is a little surprising for the first reference to an object.

3 Motivation

While the results in the TUNA challenge (and also in some other less extensive challenges) were quite satisfactory, these systems have the essential drawback of being dependent on that corpus. Similarly, this assessment also holds for the principled approaches referrred to in the last paragraph of the previous section, since they do not attempt to generalize over the corpus examined, which is even the case for the study by Viethen, Dale and Guhe (2011b). Altogether, abstracting from some given corpus is crucial, since it is unrealistic to make a new corpus evaluation for each application, corpora

being rare and typically small, if available at all. In order to increase the generality of the corpus interpretation, the results must be lifted to more general grounds, so that they can be transferred to other, somehow similar domains.

Unfortunately, learning algorithms and the structure of their results are hardly useful for this purpose – they are widely human-inaccessible, without connection to easily understandable conceptions, and a comparison of results across corpora and domains is hard to imagine. In order to enable reasonable comparisons, we attempt to formulate regularities over attributes of objects and situational properties that can be tested against a corpus. In order for a regularity to qualify for this purpose, we require that it must be

- expressed in cognitive meaningful terms,
- as domain-independent as possible, and
- hold over the entire corpus to a significant degree.

The hope is that these regularities can reasonably be transferred to related domains by accommodating the domain- and corpus-dependent parts, since these regularities contain a reasonable share of domain-independent factors that can be transferred with little adaptation. Since regularities of this kind always contain some degree of domain- and/or corpus dependency, abstraction is a crucial conponent in the formulation of the regularity or in expressing the transfer method.

A major source for our motivation is the observation that previous approaches do not take the proper task, which is *identification* of objects, into account. Their corpus analyses would also work if the purpose of the expressions in the corpus would be descriptions of properties that the subjects like or dislike or have some other attitude against. We are convinced that the task to accomplish, identification, has some, possibly an essential influence on the choice of attributes. It must make a difference whether the task is even easier than average – e.g., if one salient attribute is sufficient for achieving identification – or whether producing an identifying description is really challenging – e.g., if several attributes are needed for obtaining identification, including some less salient ones.

4 Hypothesizing Regularities

Our basic idea is to establish relations between the properties of the situation in which the subjects have chosen some expressions and properties of these expressions, and to aggregate over these relations for similar situations, to find commonalities among the association between given situations and expressions chosen. There are two crucial assumptions behind our approach:

- The choices made by the subjects in the creation of the corpus can be conceived in terms of components, typically by a systematic abstraction from surface expressions (this is shared by the GRE challenges).
- There are properties of the underlying situation which capture essentials in driving the subjects' choices – hence, the selection process is to a certain extent oriented on the task to be accomplished, with some personal preferences (this is not shared by systems in the GRE challenges, at least not explicitly).

Thus, we assume that people not only choose attributes on the basis of some intrinsic properties, such as salience, but also on the basis of their contribution to the identification of the intended referent. In particular, an attribute is more likely to be chosen if it alone allows the identification rather than in a situation where several objects share the value of this attribute. Depending on the contribution of attributes to the preferred expressions, the use of an attribute may be essential or of minor relevance. Also taking into account some degree of influence between attributes, we distinguish the following basic categories:

1. *obligatory* elements, that is, attributes that must be chosen in some sort of situation
2. *exclusive* alternatives, that is, two attributes where one of them but not the other must be chosen in some sort of situation
3. *optional* elements that is, attributes that may be chosen in some sort of situation
4. *contextual* factors leading to preferences in choosing among exclusive alternatives

or distinguishing situations from others where optional elements are chosen or not

In order to test whether some attribute belongs to one of these categories, aggregations over a set of situations are made; if the test is positive, then a regularity has been found which categorizes an attribute in the context of a set of situations. The set of situations which become subject to these tests are built on the basis of conceptual commonalities. This is where we incorporate properties of the task at hand: sets of situations are built in such a way that their commonality lies in how identification can be achieved. For example, in one set of situations identification may be possible by a single attribute, in another set of situations, a pair of attributes is required. A further distinction is whether the attribute to be examined belongs to a set of attributes that represents a minimally distinguishing description, or whether it is some extra, typically salient attribute.

We do not expect to find regularities that provide one hundred percent agreement about the use of some element in preferred expressions. Moreover, corpus data can be noisy, since humans are inherently fallible. We do, however, expect sets of observations that qualify as regularities to hold over a significantly large subset.

These regularities are interpreted as a set of rules, which are intended as a backbone of a procedure that performs the same task as the subjects in the controled experiments. It is hoped that these rules capture essentials of the rationale underlying the choices made in a better way than mere surface frequencies. Then the rules can be used in principled selection procedures, hopefully even beyond the scope of the given corpus. The success of our method depends on two crucial factors:

- the identification of properties which have a chance of leading to useful discriminations
- carefully selecting and efficiently organizing the aggregation over sets of situations that enables one to test whether or not the properties suspected to lead to good discriminations indeed do so

In what follows, we present the formalization of these issues.

5 Formalization

In formal terms, a situation is conceived as a set of properties expressed as attribute value pairs, $S = \{(a_1,v_1),\ldots,(a_n,v_n)\}$, and the result as a set of components $R = \{e_1,\ldots,e_n\}$. In a pair (a,v), a is an attribute or a predicate about the attribute's contribution to the identification (prototypically, *distinguishing*), and v is a value resp. a subset of attributes of the intended referent. e is either an attribute (implicitly including all values), or a specific attribute-value pair. A trial, that is, an individually identifiable piece in the corpus, is then an association between a situation S (only with the attribute-value pair variant) and a result R, represented as $T = (S,R)$.

Aggregations of trials are formed over common properties of the situations in these trials (with a predicate about the contribution to identification), so that a set of trials $ST = \{(S_1,R_1),\ldots,(S_n,R_n)\}$ such that a set of attribute value pairs $CP=\{(a_1,v_1), \ldots,(a_n,v_n)\}$ is common to all situations: $\forall i=1,n$: $S_i \supset CP$.

In order for a regularity to fulfil the requirements of a conceptual relation stated above, the following constraints must hold, correspondingly:

1. *obligatory elements* e_{obl}
   e_{obl} must occur in the results of ST in most cases, at least as often as $thresh_{obl}$

   $e_{obl} \in R_i$ for some i: $|(S_i,R_i)| / |ST| > thresh_{obl}$

2. *exclusive alternatives* e_{alt1}, e_{alt2}
   either e_{alt1} or e_{alt2} must occur in most of the results of ST, at least as often as $thresh_{alt1}$, each of them in several, at least as often as $thresh_{alt2}$, while they generally do not co-occur, with exceptions less than $thresh_{alt3}$

   $e_{alt1} \in R_i$ for some i, $e_{alt2} \in R_j$ for some j:
   $|(S_i,R_i) \cup (S_j,R_j)| / |ST| > thresh_{alt1}$ ^
   $|(S_i,R_i)| / |ST|$, $|(S_j,R_j)| / |ST| > thresh_{alt2}$ ^
   $|(S_i,R_i) \cap (S_j,R_j)| / |ST| < thresh_{alt3}$

3. *optional elements* e_{opt}
   e_{opt} must occur in the results of ST in some cases, at least as frequent as $thresh_{opt}$, but it must not be obligatory and it must also not appear in a pair of exclusive alternatives (second part omitted in the formalization)

   $e_{opt} \in R_i$ for some i: $|(S_i,R_i)| / |ST| > thresh_{opt}$

Situations S_1		Results	Situations S_2		Results
(a_1,v_1)	(a_2,v_3)	$\{e_1,e_2\}$	(a_1,v_2)	(a_2,v_3)	$\{e_1,e_2,e_4\}$
(a_1,v_1)	(a_2,v_3)	$\{e_1,e_3\}$	(a_1,v_2)	(a_2,v_3)	$\{e_1,e_2\}$
(a_1,v_1)	(a_2,v_3)	$\{e_1,e_2,e_4\}$	(a_1,v_2)	(a_2,v_3)	$\{e_1,e_2\}$
(a_1,v_1)	(a_2,v_3)	$\{e_1,e_3\}$	(a_1,v_2)	(a_2,v_3)	$\{e_1,e_2\}$
(a_1,v_1)	(a_2,v_3)	$\{e_1,e_2\}$	(a_1,v_2)	(a_2,v_3)	$\{e_1,e_2,e_4\}$
(a_1,v_1)	(a_2,v_4)	$\{e_1,e_2,e_4\}$	(a_1,v_2)	(a_2,v_4)	$\{e_1,e_3\}$
(a_1,v_1)	(a_2,v_4)	$\{e_1,e_3,e_4\}$	(a_1,v_2)	(a_2,v_4)	$\{e_1,e_3\}$
(a_1,v_1)	(a_2,v_4)	$\{e_1,e_2\}$	(a_1,v_2)	(a_2,v_4)	$\{e_1,e_3,e_4\}$
(a_1,v_1)	(a_2,v_4)	$\{e_1,e_3\}$	(a_1,v_2)	(a_2,v_4)	$\{e_1,e_3\}$
(a_1,v_1)	(a_2,v_4)	$\{e_2,e_3\}$	(a_1,v_2)	(a_2,v_4)	$\{e_2,e_3\}$

Table 1. Illustrating categories of components

4. *contextual factors* (a_{cf},v_{cf})
   A contextual factor (a_{cf},v_{cf}) that is considered the driving force behind the choice among exclusive alternatives e_{alt1} and e_{alt2}, in the sense that it appears in the situations where one of the exclusive alternatives is part of the chosen expression, while it does not appear in the situations where the other exclusive alternative is part of the chosen expression, with exceptions less than $thresh_{cf}$

   $e_{alt1} \in R_i$ for some i, $e_{alt2} \in R_j$ for some j:
   $\forall k$: $(a_{cf},v_{cf}) \in S_k$: $|S_k \cap S_i| > thresh_{cf} \wedge$
   $|S_k \cap S_j| < (1 - thresh_{cf})$

Table 1 illustrates these categories of elements. There are two sets of situations, S_1 on the left half, and S_2 on the right, with their associated results. (a_1,v_1) is the property common to S_1, (a_1,v_2) the one common to S_2. e_1 is an obligatory element in S_1 and S_2 with $thresh_{obl} \leq 0.9$. e_2 and e_3 are exclusive alternatives in S_1 and S_2 ($thresh_{alt1} \leq 1.0$, $thresh_{alt2} \leq 0.4$, $thresh_{alt3} \geq 0.1$), even combined with a contextual factor in S_2 ($thresh_{obl} \leq 0.9$). e_4 is an optional element ($thresh_{opt} \leq 0.3$).

The thresholds in this example are purely the result of calculations based on the data, that is, they correspond precisely to the number of cases that fulfil the respective predicates – we have chosen ten instances to make the computations simple. An independent question is, how reasonable thresholds can be nailed down in numerical values. We think that the values in the example are plausible ones, but it is not clear how much weaker they may get – for example, a threshold of around 0.6 may build a transition between an obligatory and an optional element. More practical corpus examinations are needed.

6 Preliminary Results

We have applied our method to the publically available segment of the TUNA corpus. The corpus comprises referring expressions from two domains: the identification of a piece of furniture resp. a person out of a set of such items, presented in a small grid (v. d. Sluis, Gatt, and van Deemter 2006). In the furniture domain, attributes include the type of the object, its color, size, and orientedness. In the people domain, attributes most used are beardedness, wearing glasses, age, hair, and its color. In both domains, the positions on the grid are attributes. The result is simply the subset of attribute-value pairs attributed to the intended referent in the referring expression chosen by the subjects.

In addition to that, we have enhanced the representation of situations by several attributes that we thought might be driving forces in the selection of attributes for the referring expression. The ones we have built and tested so far are essentially based on two concepts:

1) subcategories of attributes (an example of linguistic evidence), and
2) contribution to identification of an object (an example of a task-specific property).

Subcategorization comprises
1) the type,
2) most salient attributes (here: color, beardedness, wearing glasses),
3) location, and
4) remaining attributes.

Concerning the contribution to identification, we distinguish for an attribute whether
1) it allows identification by itself,
2) does so together with the type attribute,
3) does so in connection with the type and a most salient attribute, and
4) neither of these.

Hence, these distinctions allow one to discriminate between varying complexities of the underlying identification task.

Based on these attributes, we have selectively tested a number of aggregations, the set of similar trials presented to subjects, which differ only in the positions of the items on the grid, and some further aggregations, combining sets of trials with comparable task complexity

according to the measure introduced above. Within these aggregations, we have examined several attribute-value pairs in the set of results, as to whether their uses qualify for one of the regularities as defined in the previous section. Specifically, we have tested the role of the most salient attributes color, wearing glasses, and beardedness, we have made a comparison between size and orientation of pieces of furniture, and we have tested the role of some values of a person's hair (color, no hair).

The results are listed in Table 2. This Table contains the attributes that categorize the set of situations aggregated and the regularity derived for each set of situations. In the furniture domain, color was always used very often (regularity 1). If orientation resp. size gives a distinguishing description together with type and color, orientation resp. size and location are conceived as alternatives (regularities 2 and 3). Having a beard is at least optional (regularity 4), but obligatory if it is distinguishing (regularity 5). Similar regularities are derived for wearing glasses and hair color. Finally, hair color, if distinguishing, is conceived as an alternative to location (regularity 6). All thresholds involved are at least .75 ($thresh_{oblig}$), resp. .33 ($thresh_{alt2}$ and $thresh_{aalt}$). We did not discover any contextual factors; preferences for the use of position attributes can be grounded in personal choices (Bohnet 2008), but we did not model this aspect. Our major findings include the better effectiveness of color of pieces of furniture (*obligatory*) than color of hair (only *exclusive* alternative), and more frequent uses of position with increasing task complexity.

Set of situations	*Regularity*
1. furniture domain	obligatory (color)
2. distinguishing (type+color+orientation)	alternatives (position,orientation)
3. distinguishing (type+color+size)	alternatives (position,size)
4. applicable (beardedness)	optional (beardedness)
5. distinguishing (beardedness)	obligatory (beardedness)
6. distinguishing (hair color)	alternatives (hair color,position)

Table 2. Regularities found for the GRE task

6 Discussion

The regularities found can form the backbone of a choice mechanism in an NL generation component – obligatory elements are collected, one out of each set of the exclusive alternatives is taken, and optional elements are added until a distinguishing description is obtained. Choices in this procedure can be made more specific by the corpus frequencies, thus incorporating some element of the majority of approaches to the GRE challenge (such as (Bohnet 2008) and (Kelleher, McNamee 2008)). In contrast to these approaches, which are strictly performance-oriented, we envision a distribution of forces between human modeling of linguistically motivated and task-relevant factors and computation of the role of these factors regarding the choice among alternatives. In addition, some representation elements, notably aggregations and exclusive alternatives, give us more expressiveness than mere frequencies. As a result, we obtain a set of pieces of symbolic knowledge, which increase understanding of the task and are likely to pertain beyond the given corpus to some extent.

The regularities found constitute a set of crisp and cognitively meaningful rules; to some extent, they encapsulate particularities of the corpus against which they were tested. In terms of specificity, they are more concrete and detailed than principles tested on the basis of controled experiments. Conversely, these regularities are less specific than results obtained by learning methods.

A crucial question is to what extent our results can be accommodated for transferring regularities to related domains, and what data is missing for that purpose. The two domains examined, people and furniture, are significantly different from one another to discuss possible cross-relations, with the only commonality in terms of the grid, that is, location attributes. A comparison of regularities between the two domains shows that impacts of the domain-independent factors, that is, the cardinality of a minimally identifying expression, and the domain-specific properties, that is, the attributes, are interwoven. For example, *color*, a seamingly salient attribute in the furniture domain, is *obligatory* over the whole corpus, while a salient attribute in the people domain, *beardedness*, is only *optional*, unless it is *distinguishing* by itself. May be, this is an

impact of the presence of another very salient attribute, *wearing glasses*; in the furniture domain, *color* stands out in terms of salience. Moreover, the role of *hair color*, which might be considered as ontologically related to *color* in the furniture domain, is much less prominent than *color*: even in cases where it is *distinguishing* by itself, it is only *alternative* to *location*. However, this result may be an impact of the pictures used in the experiments: they all showed scientists, and one might suspect that the role of hair color would be more prominent in other kinds of situations, e.g., for identifying attractive women.

These observations suggest a number of extensions and further uses. First of all, applying our method to a larger set of corpora would not only extent the coverage beyond people and pieces of furniture, but it would also enable different views on these kind of entities in varying situations and salience. For example, a significantly increased examination of the role of attributes and their combinations might then be possible, which is inhibited by data sparseness in the TUNA corpus and also by the fact that the corpus appears to be biased in some ways. For example, there are plenty of instances where beardedness or wearing glasses are distinguishing attributes by themselves, but this is not the case for most other attributes (e.g., wearing a tie). In addition to the increased quantity of data, it is necessary to make more fine-grained distinctions of salience categories than we did so far. In particular, a context-dependent aspect appears to be useful, which would allow one to distinguish attributes that stand out in terms of salience (such as color in the furniture domain) from similarly salient attributes – there exist several in the TUNA corpus (such as wearing glasses and beardedness). As a consequence, the number and complexity of regularities would increase.

Our general idea is that a transfer to other domains looks promising on the level of some sort of salience categories; the success of this method relies on the following assumptions:

1) people behave similarly in comparable situations (easy or difficult identification task)
2) people behave similarly in comparable perception circumstances (attribute salience)
3) salience can be reasonably generalized across situations and domains

Provided these assumptions hold, a big gain can be achieved, since assessing salience categories in some other domain appears to be much less costly than creating a corpus; moreover, such assessments may serve also other purposes than GRE. Furthermore, regularities with references to attributes abstracted into salience categories are entirely domain-independent and ready for transfer, that is, to be instantiated by attributes of suitable salience categories in the target domain.

Altogether, the results are unlikely to get as accurate as this can be done by the use of learning procedures. However, if transferring is working reasonably well to domains where learning methods are not applicable - due to lack of corpora, we can potentially achieve a big gain: decision criteria are grounded in abstractions from empirical data, which is superior to using hand-crafted rules.

7 Conclusion and Further Work

In this paper, we have presented a method for finding out relations between task-relevant situational properties and components of the expressions used in a corpus that features human preferences in the GRE subtask. We have described an application to the TUNA corpus, which uncovered some yet unobserved regularities of language use in this corpus. Since the criteria used in our method are reasonably general, we believe that some of our findings also pertain beyond the TUNA corpus and even beyond its domains.

There are at least three directions for further extensions of our approach. An obvious one is the application to other corpora in the GRE task. Another direction concerns methodological improvements – so far, choosing and testing suitable aggregations has been done semi-automatically; in the long run, this should be done by a fully automated procedure. Finally, we expect that these directions of extensions will suggest refinements in the description of regularities, e.g., more than two exclusive alternatives, and some more complex dependencies may need to be modeled, especially more fine-grained situational contexts for *optionals*.

References

Belz, A., and Gatt, A. 2007. The Attribute Selection for GRE Challenge: Overview and Evaluation Results. In Proceedings of the *Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, pp. 75–83, Copenhagen, Denmark.

Bohnet, B. 2008. The Fingerprint of Human Referring Expressions and their Surface Realization with Graph Transducers (IS-FP, IS-GT, IS-FP-GT). In Proceedings of the *5th International Natural Language Generation Conference (INLG'08)*, pp.104-112, Salt Fork OH, USA.

Bohnet, B., and Dale, R. 2005. Viewing Referring Expression Generation as a Search Problem. In Proceedings of the *19th International Joint Conference on Artificial Intelligence (IJCAI-2005)*, pp. 1004-1009, Edinburgh, Scotland.

Dale, R. 1989. Cooking up Referring Expressions. In Proceedings of the *27th Annual Meeting of the ACL*, pp. 68-75.

Dale, R., and Reiter, E. 1995. Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science* 18, pp. 233-263.

Gatt, A., and Belz, A. 2008. Attribute Selection for Referring Expression Generation: New Algorithms and Evaluation Methods. In Proceedings of the *5th International Natural Language Generation Conference (INLG'08)*, pp. 50-58, Salt Fork OH, USA.

Gatt, A. and Belz, A. 2010. Introducing shared task evaluation to NLG: The TUNA shared task evaluation challenges. In Emiel Krahmer and Mariet Theune, editors, Empirical Methods in Natural Language Generation. Springer Verlag, Berlin, pp. 264–293.

Gatt, A., Belz, A., and Kow, E. 2008. The TUNA Challenge 2008: Overview and Evaluation Results. In Proceedings of the *5th International Conference on Natural Language Generation (INLG'08)*, pp. 198–206, Salt Fork OH, USA.

Gatt, A., Belz, A., and Kow, E. 2009. The TUNA–REG Challenge 2009: Overview and Evaluation Results. In Proceedings of the *12th European Workshop on Natural Language Generation (ENLG-09)*, pp. 174–182, Athens, Greece.

Gatt, A, van der Sluis, I., and van Deemter, K. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In Proceedings of the *11th European Workshop on Natural Language Generation (ENLG-07)*, pp. 49–56, Schloss Dagstuhl, Germany.

Guhe, M. and Bard, E. 2008. Adapting referring expressions to the task environment. In Proceedings of the *30th Annual Conference of the Cognitive Science Society (CogSci)*, pages 2404–2409, Austin, TX.

Gupta, S., and 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., and Walker, M. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24, pp. 157–194.

Kelleher, J., and MacNamee, B. 2008. Referring Expression Generation Challenge 2008 DIT System Descriptions (DIT-FBI, DIT-TVAS, DIT-CBSR, DIT-RBR, DIT-FBI-CBSR, DIT-TVAS-RBR). In Proceedings of the *5th International Natural Language Generation Conference (INLG'08)*, Salt Fork OH, USA.

Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Moore, J., and Oberlander, J. 2010. The first challenge on generating instructions in virtual environments. In Emiel Krahmer and Mariet Theune, editors, Empirical Methods in Natural Language Generation. Springer Verlag, Berlin, pp. 328–352.

Krahmer, E. and van Deemter, K. 2012. Computational Generation of Referring Expressions: A Survey. *Computational Linguistics* 38(1), pp. 173-218.

Pechmann, T. 1989. Incremental speech production and referential overspecification. *Linguistics* 27, pp. 98–110.

van Deemter, K., Gatt, A., van der Sluis, I., and Power, R. 2012. Generation of referring expressions: Assessing the Incremental Algorithm. *Cognitive Science*, 36(5) pp. 799-836.

van der Sluis, I., Gatt, A. and van Deemter, K. 2006. Manual for the TUNA Corpus: Referring expressions in two domains. Technical Report AUCS/ TR0705, University of Aberdeen.

Viethen, J., and Dale, R. 2008. The use of spatial relations in referring expression generation. In Proceedings of the *5th International Conference on Natural Language Generation (INLG'08)*, pp. 59–67, Salt Fork OH, USA.

Viethen, J., and Dale, R. 2011. GRE3D7: A Corpus of Distinguishing Descriptions for Objects in Visual Scenes In Proceedings of the *UCNLG+Eval: Language Generation and Evaluation Workshop*, pp. 12–22, Edinburgh, Scotland, UK.

Viethen, J., Dale, R., and Guhe, M. 2011a. The Impact of Visual Context on the Content of Referring Expressions. In Proceedings of the *13th European Workshop on Natural Language Generation (ENLG-11)*, pp. 44–52, Nancy, France.

Viethen, J., Dale, R., and Guhe, M. 2011b. Generating subsequent reference in shared visual scenes: Computation vs. re-use. In Proceeding of the *2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1-9, Edinburgh, UK.

Viethen, J., Zwarts, S., Dale, R., and Guhe, M. 2010. Dialogue reference in a visual domain. In Proceedings of the *7th International Conference on Language Resources and Evaluation (LREC)*, pp. 1-1, Valetta, Malta.