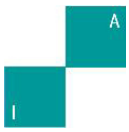
Inteligencia Artificial 45(2010), 48-58
doi: 10.4114/ia.v14i45.1100



INTELIGENCIA ARTIFICIAL

http://erevista.aepia.org/

From Semantic Properties to Surface Text: the Generation of Domain Object Descriptions

Diego Jesus de Lucena, Daniel Bastos Pereira, Ivandré Paraboni
School or Arts, Sciences and Humanities, University of São Paulo
Av. Arlindo Bettio, 1000 - São Paulo (Brazil)
{diego.si,daniel.bastos,ivandre}@usp.br

Abstract At both semantic and syntactic levels, the generation of referring expressions (*REG*) involves far more than simply producing 'correct' output strings and, accordingly, remains central to the study and development of Natural Language Generation (*NLG*) systems. In particular, *REG* algorithms have to pay regard to humanlikeness, an issue that lies at the very heart of the classic definition of Artificial Intelligence as, e.g., motivated by the Turing test. In this work we present an end-to-end approach to *REG* that takes humanlikeness into account, addressing both the issues of semantic content determination and surface realisation as natural language.

Keywords: Natural Language Generation, Referring Expressions, Attribute Selection, Surface Realisation.

1 Introduction

When talking about objects or entities in a particular domain, human speakers make extensive use of *referring expressions* such as 'this article', 'They', 'the man next door', 'Queen Victoria', 'a wooden house' and so on. Much more than simply providing written or spoken discourse labels, however, referring expressions are to a great extent responsible for the very cohesion and coherence of the discourse. The choices made by human speakers at both syntactic and semantic levels of reference actually tie discourse units together, and may determine whether the underlying structure is fluent, or even whether it makes any sense at all. For these reasons, and also due to a number of nontrivial computational challenges involved, the generation of referring expressions (*REG*) remains central to the study and development of Natural Language Generation (*NLG*) systems to date [18], and it is the focus of this paper as well[1].

Although at first it may seem as a narrow research field, *REG* involves far more than simply producing correct (e.g., unambiguous, well-formed) output strings, comprising two closely-related but distinct research problems: the selection of the appropriate semantic properties to be included in the output description (called the 'attribute selection' task) and the choice of the appropriate wording (the 'surface realisation' task.) Remarkably, the notion of appropriateness - or humanlikeness - permeates both issues, and (not unlike many other aspects of *NLG*), referring expressions are required, at the very least, to appear *plausible* from the psycholinguistic point of view in both semantics and surface form.

Consider the following examples. First, given the need to refer to a certain domain object (e.g., a person) with known semantic properties (e.g., the property of BEING BLONDE and BEING TALL), will a human speaker say 'the tall person' or 'the blonde person'? Second, even if we know in advance which

[1]We actually focus on a particular kind of *NLG* application, namely, which generates text from (usually) non-linguistic input data.

ISSN: 1988-3064(on-line)

properties should be used in that particular reference (e.g., the property of BEING A FEMALE), will we say 'the girl' or 'the woman'? The difference among these alternatives may be subtle, but it is nevertheless the key to the design of systems that generate natural language in the same way as we do.

Unlike much of the existing work in the field, in this paper we will address both issues of attribute selection and surface realisation, presenting an end-to-end approach to *REG* (i.e., from semantic properties to words) that takes humanlikeness into account. In doing so, we shall focus on the generation of instances of *definite descriptions* such as those ubiquitously found in human-computer interaction, virtual environments and applications dealing with visual and/or spatial domains in general (e.g., 'Please press *the green button*'), assuming that the choice for this particular linguistic form has already been made, as opposed to, e.g., the use of a pronoun as in 'Please press *it*' or a proper name as in 'Call *Mr. Jones*'.

The reminder of this paper is structured as follows. Section 2 describes the data set that will be taken as the gold standard for *REG* in both attribute selection and surface realisation tasks. The tasks themselves are described individually in Section 3 (attribute selection) and Section 4 (surface realisation.) Finally, Section 5 presents the results of our approach and Section 6 draws our conclusions.

2 Reference Data

The computational generation of referring expressions that take humanlikeness into account poses the question of how to measure closeness to human performance. In this work we will use the *TUNA* corpus of referring expressions presented in [8, 20] as a gold standard for both attribute selection and surface realisation tasks. *TUNA* is a database of situations of reference collected for research and development of referring expressions generation algorithms. Each situation of reference (called a *TUNA trial*) comprises a number of referable objects represented by sets of semantic properties. One particular object is the *target object* (i.e., the intended referent in that trial) and all the others are simply *distractor objects* (i.e., competing objects found in the reference context).

In addition to a target and its distractors, each situation of reference in *TUNA* includes an unambiguous *description* of the target object as produced by a native or fluent speaker of English, participant of a controlled experiment. Each description conveys a set of semantic properties selected by a human speaker accompanied by its surface string (i.e., the actual words uttered by each participant.) Put together, the collection of discourse objects and description (represented both at semantic and surface levels) provides us with all the required information for both attribute selection and surface realisation of referring expressions as those produced by human speakers, i.e., taking the issue of humanlikeness into account.

TUNA descriptions are available in two domains: FURNITURE (indoors scenes conveying pieces of furniture such as sofas, chairs etc.) and PEOPLE (human photographs.) All domain objects and descriptions are uniquely identifying sets of semantic properties represented as attribute-value pairs in *XML* format.

Domain-specific attributes include TYPE, COLOUR, SIZE, ORIENTATION, AGE, HAIR COLOUR and so on. In addition to these, in some trials the participants were encouraged to use the attributes X-DIMENSION or Y-DIMENSION as well, representing the relative position of the objects within a 3 x 5 grid on the computer screen of the experiment. In the *TUNA* corpus this trial condition is marked by the tags *LOC−* and *LOC+*. The following is an example of a semantic specification in *TUNA*, adapted from [8, 20]:

```
<DESCRIPTION>
  <ATTRIB ID="a1" NAME="size" VALUE="large"/>
  <ATTRIB ID="a2" NAME="colour" VALUE="red"/>
  <ATTRIB ID="a3" NAME="type" VALUE="chair"/>
  <ATTRIB ID="a4" NAME="y-dimension" VALUE="1"/>
  <ATTRIB ID="a5" NAME="x-dimension" VALUE="2"/>
</DESCRIPTION>
```

In the attribute selection task we will use the entire set of 720 instances of singular reference available from the *TUNA* corpus, comprising 420 instances of Furniture and 360 instances of People descriptions. In the surface realisation task, given that we intend to translate the existing descriptions to Portuguese as discussed later, we will focus on a subset of 319 singular descriptions on the Furniture domain only. Each of these tasks are discussed in turn in the next sections.

3 Attribute Selection

In this section we will focus on the task of determining the semantic content of referring expressions, i.e., the task of deciding *what* to say, known as the attribute selection (*AS*) task of referring expressions generation. Linguistic realisation issues (i.e., *how* to say it) will be dealt with in Section 4.

3.1 Background

The computational task of attribute selection (*AS*) has received a great deal of attention in the *NLG* field for nearly two decades now, e.g., [4, 10, 11, 16, 5], and has been the subject of a number of recent *NLG* competitions, e.g., [1]. The core aspects of the *AS* problem have been established in [4], in what is probably the most influential wok in the field to date, called the *Incremental algorithm*. In this approach domain objects are represented as sets of semantic properties, i.e., attribute-value pairs as in (COLOUR, black), and always include a TYPE attribute representing the property normally realised as the head noun of a definite description. For example, (TYPE, cube) may be realised as 'the cube'.

The input to the algorithm is a target object r (i.e., the entity to be described), a context set C containing the distractors of r (i.e., the set of objects to which the reader or hearer is currently attending, and from which r has to be distinguished) and the semantic properties of each object. The algorithm makes use also of a list of preferred attributes P to specify the order of preference in which attributes should be considered for inclusion in the description under generation. For example, in a particular domain the list P may determine that the generated descriptions should preferably convey the attributes $P = <$TYPE, SIZE, COLOUR$>$, in that order.

The output of the algorithm is a set of properties L of the target object r, such that L distinguishes r from all distractors in the context set C. The resulting set of attributes corresponds to the semantic contents of a description of r, and it may (subsequently) be realised, for example, as a definite description in natural language.

The set L is built by selecting attributes that denote r but which do not apply to at least one of its distractors, in which case the property is said to have *discriminatory power* or to *rule out* distractors. A set of attributes that rules out all the distractors in C comprises a *uniquely distinguishing* description of r. Consider the following example of four domain objects - three cubes and a cone of various sizes and colours - and their referable properties.

```
Obj1: (type, cube), (size, small), (colour, black)
Obj2: (type, cube), (size, large), (colour, white)
Obj3: (type, cone), (size, small), (colour, black)
Obj4: (type, cube), (size, small), (colour, white)
```

The Incremental algorithm works as follows. Assuming the order of preference $P = <$TYPE, SIZE, COLOUR$>$, a description of *Obj*1 may be represented by the set of properties $L =$ ((TYPE, cube), (SIZE, small), (COLOUR, black)), which then could be realised as 'the small black cube'. The use of the first property (i.e., the property of BEING A CUBE) rules out *Obj*3 (which is a cone); the use of the second property (i.e., the property of BEING SMALL) rules out *Obj*2 (which, despite being a cube, is not small but large) and the use of the third property (i.e., the black colour) rules out *Obj*4 (which is white.) Analogously, a description of *Obj*2 could be realised as 'the large cube', *Obj*3 could be described simply as 'the cone', and *Obj*4 could be described as 'the small white cube'[2].

The work in [4] attempts to avoid the inclusion of properties that do not help ruling out distractors by selecting only those attributes that have some discriminatory power, and by finalizing as soon as a uniquely distinguishing attribute set is found. The reason for favouring short descriptions in this way is the recognised risk of producing false conversational implicatures in the sense defined by H. P. Grice [9]. For instance, in a context in which there is only one object of type 'cube', a logically redundant reference to the colour as in 'Please move the *red* cube' may force the hearer to wonder why this attribute has been mentioned at all, or even whether a second cube (presumably of a different colour) remains unnoticed in the context. In this case, a short, non-redundant description as, e.g., 'the cube' would have

[2]For simplicity, semantic properties in our examples will match individual English words (e.g., (TYPE, cone) is realized as 'the cone'), but this does not have to be so. For instance, a single property of (a box) 'having been bought in 1952' may be variously realised as 'the box *bought in 1952*', 'the box *bought in the early fifties*', 'the *old* box', and so on.

been much preferred. On the other hand, given that the computational task of finding minimal attribute sets is known to be a NP-hard problem [4], in the Incremental approach once an attribute is selected, it can never be removed, i.e., the algorithm does not backtrack even if a subsequent inclusion renders a previously selected attribute redundant (hence the name 'Incremental'.)

The Incremental algorithm remains the basis of many (or most) *AS* algorithms to date, although more sophisticated instances of reference phenomena have been addressed in many of its extensions[3]. At the most basic level, *AS* algorithms in general are required to guarantee *uniqueness*, i.e., producing uniquely identifying descriptions that denote the target object and no other distractor in the given context. However, it is often the case that *AS* algorithms are required to take *humanlikeness* into account as well (i.e., ideally selecting those same semantic properties that a human speaker would have selected in that context) whilst paying regard to *brevity* (i.e., avoiding the generation of overly long or otherwise clumsy descriptions that may lead to false implicatures.)

Which precise factors - uniqueness, humanlikeness, brevity and others - may be favoured by a particular *AS* algorithm based on the Incremental approach is greatly dependent on how the list P of preferred attributes is defined. More specifically, by changing the order (or the attributes themselves) in P, the behaviour of the *AS* strategy may change dramatically. For example, had we defined the list as $P =$ <COLOUR, SIZE, TYPE> in the previous example, then *Obj*3 would have been described as 'the small black cone', and not simply as 'the cone' as one could obtain for $P =$ <TYPE, SIZE, COLOUR>.

Dale and Reiter [4] do not provide details on how exactly the ordering of attributes in P should be defined to meet these criteria, but they do suggest that P should be worked out from the domain. Thus, a first alternative would be to order the list P according to the *discriminatory power* of the existing attributes, i.e., by selecting first the attribute capable of ruling out *the largest possible number of distractors*. This strategy minimizes the risk of false implicatures, and for that reason it is implemented by many *Greedy* or *Minimal* approaches to *AS*. However, if minimal descriptions may be desirable in some applications, they may simply look unnatural in many others. For instance, a minimal description of *Obj*2 in the previous example would be realised simply as 'the large (object)'.

Instead of favouring highly discriminating attributes, another possible way or ordering P is by selecting first the attributes *most commonly seen* in the domain. Descriptions produced in this way tend to be longer than necessary to avoid ambiguity, but may also seem closer to real language use and, arguably, more 'human-like'. Brevity and humanlikeness may therefore be conflicting goals, and a balance between the two is called for.

3.2 Current Work

We intend to develop an *AS* algorithm primarily focused on humanlikeness, that is, an algorithms that selects sets of 'typical' attributes as close as possible to what human speakers would produce, but which pays (to a lesser extent) regard to brevity as well. To this end, we will simplify and assume that 'typical' attributes are those found *most frequently* in the domain, and we will derive a general *AS* strategy in which the attributes in P are selected in descending order of *relative frequency* as seen in a collection of definite descriptions (in our case, the *TUNA* corpus.) Besides implementing the policy of selecting typical attributes first, this will allow the inclusion of highly frequent attributes such as TYPE (usually required to form a head noun) to be modelled in a more convenient, domain-independent way than in the Incremental approach. The most frequent attributes in both FURNITURE (left) and PEOPLE (right) domains are shown in Table 1. Our general *AS* strategy will use the frequency lists in Table 1 with one important exception: as mentioned in Section 2, there are two types of *TUNA* trials: one kind (marked by the tag *LOC*+) in which the participants were encouraged to make reference to the screen coordinates (i.e., using the X- and Y-DIMENSION attributes), and a second type in which the participants were told to avoid these attributes (*LOC*−). Thus, likewise [7] we will assign maximum priority to the X- and Y-DIMENSION attributes in *LOC*+ trials and, conversely, in *LOC*− situations we will keep these attributes in their relative (frequency-based) position in P.

Favouring the most frequent attributes gives rise to the question of whether some attributes could be so

[3]For instance, as the context becomes more complex, there is a natural trend towards the use of relational properties as well (e.g., [11]), as in 'the cube next to the small cone', and this may be the case even if the reference to the cone is redundant from the logical point of view (e.g., [16]).

Furniture		People	
Attribute	Freq.	Attribute	Freq.
type	97.18%	type	92.70%
colour	86.52%	hasBeard	44.89%
size	36.99%	hasGlasses	42.70%
orientation	33.23%	y-dimension	31.02%

Table 1: Most frequent attributes in the corpus

common as to be *always* selected even when they do not help ruling distractors. Indeed, in both domains in Table 1 we observe a significant drop in frequency after the first few instances, making a clear divide between highly frequent attributes and the reminder. This may suggest a simple AS strategy that selects all attributes whose frequency falls *above a certain threshold value* v regardless of their discriminatory power. To put this idea to the test, we will chose the empirical threshold value $v = 0.80$ to mark the compulsory inclusion of attributes, which in practise will grant special treatment to the attributes TYPE and COLOUR in the FURNITURE domain, and also to the attribute TYPE in the PEOPLE domain[4].

Having established the general principles of our frequency-based approach, we now turn our attention to the issue of brevity or, more precisely, to its interplay with humanlikeness. Our combined strategy can be summarised by two simple assumptions: (a) in a complex context (i.e., with a large number of objects), computing the attribute capable of ruling out the largest possible number of distractors (as in a 'greedy' or 'minimal' AS approach) may not only be hard (from the computational point of view), but also less natural than simply using 'common' attributes (as in a frequency-based approach); and (b) on the other hand, as the number of distractors decreases, it become gradually easier for the speaker to identify those attributes that are most helpful to achieve uniqueness, up to the point in which he/she may naturally switch from a frequency-based to a greedy AS strategy and finalize the description at once.

These assumptions (a) and (b) lead to the following referring expressions generation algorithm, which combines a frequency-based approach that takes location information into account (i.e., adapted to $LOC+$ and $LOC-$ situations of reference) with a greedy attribute selection strategy. The function RULESOUT ($< Ai, Vi >$) is meant to return the set of context objects for which the property $< Ai, Vi >$ is true of, and the $+/-$ signs stand for insert and remove set operations.

```
1.  L <- nil
2.  P <- preferred attributes for given LOC tag
3.  // compulsory selection above threshold frequency v
4.  for each Ai in P do
5.    if (frequency(Ai) > v)
6.       P <-  P - Ai
7.       L <-  L + <Ai,Vi>
8.       C <-  C - RulesOut(<Ai,Vi>)
9.  repeat
10. while (C) and (P)
11.    // greedy search for highly dicriminating attrib.
12.    for each Ai in P do
13.       if(RulesOut(Ai,Vi) <> C)
14.          L <- L + <Ai,Vi>
15.          return(L}
16.    // default frequency-based selection
17.    P <- P - Ai
18.    if(RulesOut(Ai,Vi)) <> nil
19.       L <- L + <Ai,Vi>
20.       C <- C - RulesOut(<Ai,Vi>)
21. repeat
22. return(L)
```

[4] Besides using a threshold value for compulsory attribute selection, we have extensively tested a number of (less successful) variations of this approach. For example, in one such test we attempted to use a list of preferred properties (i.e., attribute-value pairs) instead of preferred attributes. These variations to the current proposal were described in [6].

As in the Incremental approach, given a target object r in a context set C, the goal of the algorithm is to produce a list L of attributes such that L distinguishes r from all distractors in C. The list L is initially empty (line 1) and P is sorted in descending order of frequency; if the situation of reference is marked with $LOC+$ (which indicates that the use of location information was encouraged in that particular situation) the X-DIMENSION and Y-DIMENSION attributes will appear in the first position in P; if not, these attributes remain in their original frequency-based position.

Attribute selection proper starts with the inclusion of all attributes above the threshold value v (3-9). If a uniquely identifying description has been found, then the algorithm simply terminates (22). If not, the algorithm searches for a highly discriminating attribute Ai in the entire list P such that Ai is capable of ruling out all remaining distractors at once (13). If such Ai exists, then the algorithm terminates (15). If not, attribute selection continues by iterating through the list P and removing the next attribute Ai (line 17 - recall that P is sorted by frequency.) If Ai rules out at least one distractor in the context (18), then Ai is included in the description (19) and the corresponding distractors are removed from C (20).

The frequency-based selection (16-20) is repeated until a uniquely identifying description is obtained (that is, until the context C is empty) or until there are no attributes left in P, in which case the output description L will remain ambiguous for lack of data (10-21 loop.)

4 Surface Realisation

Having produced a set of semantic properties that uniquely describes a target object, we will now address the next step in the generation of a referring expression, i.e., the computation of a suitable linguistic description (in our case, rendered in Portuguese.) Following the same principle of 'humanlikeness as frequency' applied to the previous AS task, we will presently favour a frequency-based approach to surface realisation as well.

4.1 Background

The surface realisation of definite descriptions can be viewed as a task of producing word strings from a set of semantic properties such as those generated in the AS task in the previous section. A standard approach to this consists of writing appropriate *grammar rules* to determine the possible mappings from semantic properties to word strings, and how these strings should be combined (e.g., following agreement rules etc.) into meaningful definite description.

Alternatively, the introduction of statistical methods to *NLG* (e.g., [12, 14]) allowed the development of trainable, language-independent approaches that have been called 'generate-and-select' or '2-stages' generation. In this case, rather than using rules to determine the correct wording of the output, the generator simply makes use of a dictionary of mappings from semantics to surface forms to produce (in the so-called 'generate' stage) all possible strings from the given input, including even a (possibly large) number of ill-formed variations; in a subsequent stage (the 'select' step), a robust statistical language model selects the output string of highest probability among the available candidates.

For example, consider an input set of attributes L = ((TYPE, cube), (SIZE, small), (COLOUR, black), (X DIMENSION,1)), and a dictionary D conveying the mappings from semantic properties to multiple alternative realisations (e.g., phrases acquired from corpora) as follows:

```
(type, cube)     -> "cube";
                    "box".
(size, small)    -> "small";
                    "little".
(colour, black)  -> "black";
                    "dark".
(x-dimension, 1) -> "on the left";
                    "in the first column";
                    "in column #1".
```

Given L and D as an input, we may produce all possible permutations of the corresponding phrases or, more commonly, those that match a pre-defined template (e.g., [3], which may also be acquired from corpora) to avoid combinatory explosion. For instance, given a template in the form ('the' $size $colour

$type $x-dimension) in which the $ fields are to be filled in with phrases from the above dictionary D, the 'generate' stage would produce 2 * 2 * 2 * 3 = 24 candidate descriptions of L.

Having over-generated a set of alternative descriptions, the task of filtering out the inappropriate candidates is implemented with the aid of a language model trained from a large collection of documents in the domain under consideration. More specifically, after computing the probability p of the model generating each alternative, the string of highest p is chosen as the most likely output and the others are simply discarded. Humanlikeness in this case is once again accounted for (at least partially) by the use of the most frequent phrases from the dictionary and an adequate language model.

4.2 Current Work

We take a simple template-based statistical approach to surface realisation of definite descriptions in the FURNITURE domain as follows. First, two independent annotators made a comprehensive list of possible Portuguese text realisations represented as phrases[5] for each semantic property in the corpus. The lists were subsequently compared and merged for completeness, resulting in a set of 22 possible properties mapped into 41 phrases, including mainly prepositional (e.g., 'facing backwards'), adjectival (e.g., 'red') and noun phrases (e.g., 'chair') with their possible gender and number variations. These mappings make a dictionary structure as discussed in the previous section.

To decide which possible phrases should be combined to form output strings and how, we use a definite description template suitable to Portuguese phrase order in the form *($det $type $colour $size $orientation $x-dimension $y-dimension)* in which $det stands for the determiner of the description (i.e., a definite article) and the reminder $ slots are to be filled in with phrases from the dictionary of realisations. We notice that the template does not explicitly encode agreement rules of any kind, but simply enforces the word ordering. The definition of more linguistically-motivated templates would represent a bottleneck to our approach, requiring the development of grammar rules, or at least the use of a parsed corpus from which these rules could be inferred. Whilst these issues could be tackled fairly easily given the structural simplicity of our definite descriptions, additional work would be required to guarantee language-independency.

For a given non-linguistic input description represented as a list of semantic properties L, we compute all possible sets of phrases that match the pre-defined template description. For example, let us assume a situation in which a description comprises four semantic properties L = ((TYPE, chair), (SIZE, small), (COLOUR, red) , (ORIENTATION, backward)), in which each attribute may have two alternative realisations besides the gender variation of the adjectival phrases for the SIZE and COLOUR attributes. L in this case would be associated with 2 * 4 * 4 * 2 = 64 Portuguese phrase sets, and these would be multiplied by two once again to account for masculine/feminine determiners, making 128 possible realisations in total. Once these alternatives are (over)generated in this way, the correct (i.e., most likely) output is selected with the aid of a 40-million bigram language model from Brazilian newspapers and magazine articles.

For evaluation purposes, we will call this our *Statistical (surface realisation) System*[6]. As an alternative to this approach, and also to provide us with a (possibly strong) baseline system, we have developed a set of standard grammar rules to generate the same instances of Portuguese definite descriptions as well. These grammar rules (actually, a *DCG*) cover the entire set of definite descriptions that we intend to generate, taking gender, number and structural constraints into account. We will call this our *Rule-based (surface realisation) System*. For a comparison between the two approaches, see [19].

In both Statistical and Rule-based systems, a certain amount of errors are to be expected since the *TUNA* raw data include non-standard attribute usages (i.e., expressions that the participants were not allowed to use, but nevertheless did) which our systems will not handle. In addition to that, we presently do not combine more than one attribute into a single phrase (e.g., realising both X-DIMENSION and Y-DIMENSION attributes as in 'the upper right corner') even if the combined property usage turns out to be more frequent in the data than referring to the individual attributes. In these cases, both system will simply produce a 1-2-1 mapping (e.g., 'the 5th column in the top row') and will be penalised accordingly in the systems' evaluation.

[5] We use phrases - as opposed to words - as our smallest text unit because such phrases are fixed pieces of natural language provided directly by human annotators, and within which permutations do not have to be further considered.

[6] Further details are discussed in [17].

	Furniture		People	
Criteria	Mean	SD	Mean	SD
Dice	0.800	0.213	0.667	0.270
MASI	0.444	0.367	0.333	0.337

Table 2: Attribute Selection System results

	Furniture		People	
Criteria	Mean	SD	Mean	SD
Dice	0.768	0.240	0.634	0.241
MASI	0.536	0.367	0.322	0.271

Table 3: Attribute Selection Baseline results

5 Evaluation

In this section we will separately discuss the results of an intrinsec evaluation work applied to the attribute selection task (obtained by comparing our proposed *AS* strategy to the Dale and Reiter Incremental algorithm), and to the surface realisation task (obtained by comparing our Statistical and Rule-based generation strategies.)

5.1 Attribute Selection Results

We applied our *AS* approach to the generation of the entire set of 720 instances of singular reference available from the *TUNA* corpus, being 420 FURNITURE and 360 PEOPLE descriptions. The resulting collection of attribute sets make our *System* set, and will be compared with the same descriptions as generated by a baseline Dale and Reiter Incremental algorithm [4] as implemented in [6], which we will call *Reference* set.

The evaluation of our *AS* strategy was carried out by comparing each *System-Reference* description pair as in [1]. More specifically, we computed the Dice coefficient of similarity between sets, and its variation *MASI* (which penalises omissions more heavily.) In both cases, higher scores are better (score 1 is obtained for identical sets.) The results for both FURNITURE and PEOPLE domains are shown in Tables 2 (System set) and 3 (Baseline Reference set).

In the above we observe that our approach generally outperforms the baseline Incremental algorithm (except in *MASI* scores for FURNITURE.) Moreover, the current results (particularly for the FURNITURE domain) are also superior to those obtained by a previous version of our algorithm in [6], in which location information was not taken into account.

5.2 Surface Realisation Results

We applied both the Statistical and Rule-based approaches to surface realisation as described in Section 4.2. to generate 319 instances of singular reference in the FURNITURE domain[7]. In addition to that, we manually developed a *Reference set*[8] conveying Portuguese translations of the original (English) descriptions provided by the *TUNA* corpus. The translations were produced by two independent annotators and subsequently normalized to facilitate agreement, removing noise such as likely errors (e.g., 'red chair in center red'), meta-attribute usage (e.g., 'first picture on third row'), illegal attribute usage (e.g.., 'the grey desk *with drawers*'), differences in specificity (e.g., 'shown from the side' as a less specific alternative

[7] These instances of referring expressions were those provided as training data in [1].

[8] Given the differences between languages, and the fact that the translated descriptions were not produced in situations of real communication, our results are not directly comparable to the work done for the English language, and the present *Reference* set should be viewed simply as a standard of acceptable language use for evaluation purposes.

Criteria	Statistical	Rule-based
Edit distance	2.690	1.260
BLEU	0.377	0.800
NIST	5.442	7.022

Table 4: Surface Realisation results

to both 'facing left' and 'facing right' values) and synonymy (e.g., 'facing the viewer' as an alternative to 'facing forward'.) The *Reference* set contains Portuguese descriptions of up to 12 words in length (5.62 on average.) Further details on this data set are discussed in [17].

Provided our *System's* output and the Portuguese *Reference* set, we followed the evaluation procedure applied in [2] to compute Edit distance, *BLEU* [15] and *NIST* [13] scores for each *System-Reference* pair. Edit distance takes into account the cost of insert, delete and substitute operations required to make the generated *System* string identical to the corresponding Reference (a zero distance value indicates a perfect match.) *BLEU* - and its variation *NIST* - are widely-used evaluation metrics for Machine Translation systems, and which are useful for text evaluation in general. *BLEU/NIST* scores are intended to compute the amount of information shared between *System* and *Reference* output strings as measured by n-gram statistics. *BLEU* differs from *NIST* mainly in the way that sparse n-grams (which are considered to be more informative in *NIST* evaluation) are weighted, but both scores are known to correlate well with human judgments. For both *BLEU* and *NIST*, higher scores mean better text quality. *BLEU* scores range from 0 to 1, whereas the maximum *NIST* value depends on the size of the data set. The results are shown in Table 4.

In the comparison between the output of our systems and the *Reference* set, the main difference found was due to synonymy. For example, whereas the Rule-based or Statistical systems may have chosen the word 'line', the *Reference* set may contain, e.g., 'row', and this will be penalised accordingly by all evaluation metrics, although it may be debatable whether this actually constitutes an 'error'[9].

The results in Table 4 show that the Rule-based approach outperforms the Statistical system according to all criteria, producing descriptions that are much closer to those found in the *Reference* set. This was, in our view, to be fully expected since the grammar rules are (except for the shortcomings described in the previous section) nearly ideal in the sense that they cover most linguistic constraints expressed in the *Reference* set, and were indeed built from that very data set.

There are a number of reasons why the Statistical approach was less successful (overall, 103 instances or 32.3% were incorrectly generated in this approach), chief among them the use of a bigram language model (which is unable to handle long-distance dependencies appropriately) and the use of a template with no encoded agreement constraints. Although we presently do not seek to validate this claim, we believe that by either building a larger, more robust language model, or by enhancing the linguistic constraints in our description template, these results would most likely match those produced by the grammar rules. Moreover, although the results of the Rule-based system are presently superior to those obtained by the Statistical approach, the latter is in principle capable of generating text in any arbitrary language (i.e., as long as a sufficiently large training corpus is provided.) By comparison, the Rule-based system would require a language specialist to write new rules from scratch, which may be a costly or labour-intensive task.

6 Final Remarks

We have presented a combined approach to the generation of definite descriptions as Portuguese text that addresses both attribute selection and surface realisation tasks. Put together, these efforts constitute an

[9] Another remarkable difference was the word order of (Brazilian) Portuguese adjectives. For example, 'a large red table' could be realised either as TYPE + COLOUR + SIZE , or as TYPE + SIZE + COLOUR, and both alternatives are in principle acceptable. This may suggest that a much more sophisticated approach to Portuguese realisation is called-for, especially if compared to the generation of English descriptions, whose word order seems fairly straightforward.

end-to-end (from semantic properties to surface text) implementation of a referring expressions generation module for a possible *NLG* system.

Regarding the attribute selection task, we described a frequency-based greedy algorithm that attempts to balance brevity and humanlikeness of the generated descriptions. Our results are comparable to one of the best-known works in the field, the Dale and Reiter Incremental algorithm, and indeed closer to those produced by human speakers in two different domains as provided by a corpus of referring expressions.

With respect to surface realisation, we applied a standard 2-stage generation approach to (a) produce candidate descriptions and then (b) select the most likely output according to a bi-gram language model of Portuguese. The results were inferior to our (admittedly strong) baseline system based on grammar rules, but the comparison was still useful to reveal the weaknesses of the current approach (e.g., the need for more linguistically-motivated description templates) and also its advantages (e.g., language-independency.)

Acknowledgements

The first author has been supported by *CNPq*; the second author has been supported by the University of São Paulo, and the third author acknowledges support by *FAPESP* and *CNPq*. We are also thankful to the entire *TUNA* and *REG-2007* teams for providing us with the data used in this work.

References

[1] A. Belz and A. Gatt. The attribute selection for gre challenge: Overview and evaluation results. *Proceedings of UCNLG+MT: Language Generation and Machine Translation*, 2007.

[2] A. Belz and A. Gatt. Intrinsic vs. extrinsic evaluation measures for referring expression generation. *Proceeding of the 46th Annual Meeting of the Association for Computational Linguistics (ACL08)*, pages 197–200, 2008.

[3] I. Brugman, M. Thëune, E. Krahmer, and J. Viethen. Realizing the costs: Template-based surface realisation in the graph approach to referring expression generation. *Proceeding of the 12th European Workshop on Natural Language Generation (ENLG-2009)*, pages 183–184, 2009.

[4] R. Dale and E. Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19, 1995.

[5] R. Dale and J. Viethen. Referring expression generation through attribute-based heuristics. *Proceeding of the 12th European Workshop on Natural Language Generation (ENLG-2009)*, pages 58–65, 2009.

[6] D.J. de Lucena and I. Paraboni. Combining frequent and discriminating attributes in the generation of definite descriptions. *Lecture Notes in Artificial Intelligence*, 5290:252–261, 2008.

[7] D.J. de Lucena and I. Paraboni. Improved frequency-based greedy attribute selection. *Proceeding of the 12th European Workshop on Natural Language Generation (ENLG-2009)*, pages 189–190, 2009.

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

[9] H.P. Grice. Logic and conversation. *Syntax and Semantics*, iii: Speech Acts:41–58, 1975.

[10] E. Krahmer and M. Thëune. Efficient context-sensitive generation of referring expressions. *Information Sharing Reference and Presupposition in Language Generation and Interpretation*, pages 223–264, 2002.

[11] E. Krahmer, S. van Erk, and A. Verleg. Graph based generation of referring expressions. *Computational Linguistics*, 29(1):53–72, 2003.

[12] I. Langkilde and K. Knight. Generation that exploits corpus-based statistical knowledge. *Proceedings of COLING-ACL98*, pages 704–710, 1998.

[13] NIST. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *www.nist.gov/speech/tests/mt/doc/ngram-study.pdf*, 2002.

[14] A. Oh and A. Rudnicky. Stochastic language generation for spoken dialogue systems. *Proceedings of the ANLP-NAACL 2000 Workshop on Conversational Systems*, pages 27–32, 2000.

[15] S. Papineni, T. Roukos, W. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL02)*, pages 311–318, 2002.

[16] I. Paraboni, K. van Deemter, and J. Masthoff. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254, 2007.

[17] D.B. Pereira and I. Paraboni. Statistical surface realisation of portuguese referring expressions. *Lecture Notes in Artificial Intelligence*, 5221:383–392, 2008.

[18] E. Reiter and R. Dale. Building natural language generation systems. *Cambridge University Press*, 2000.

[19] F. M. V. Santos, D.B. Pereira, and I. Paraboni. Rule-based vs. probabilistic surface realisation of definite descriptions. *VI Workshop on Information and Human Language Technology (TIL-2008) / XIV Brazilian Symposium on Multimedia and the Web*, pages 372–374, 2008.

[20] K. van Deemter, I. van der Sluis, and A. Gatt. Building a semantically transparent corpus for the generation of referring expressions. *Proceedings of the International Natural Language Generation Conference (INLG-2006)*, 2006.