# From TUNA Attribute Sets to Portuguese Text: a First Report

**Daniel Bastos Pereira**
Escola de Artes, Ciências e Humanidades
University of São Paulo – USP
Av. Arlindo Bettio, 1000 - São Paulo, Brazil
daniel.bastos@usp.br

**Ivandré Paraboni**
Escola de Artes, Ciências e Humanidades
University of São Paulo - USP
Av. Arlindo Bettio, 1000 - São Paulo, Brazil
ivandre@usp.br

## Abstract

This document describes the development of a surface realisation component for the Portuguese language that takes advantage of the data and evaluation tools provided by the REG-2008 team. At this initial stage, our work uses simple n-gram statistics to produce descriptions in the Furniture domain, with little or no linguistic variation. Preliminary results suggest that, unlike the generation of English descriptions, contextual information may be required to account for Portuguese word order.

## 1 Introduction

In this work we describe a surface realisation component for Portuguese definite descriptions using the data and evaluation tool provided as part of the REG-2008 Challenge. However, given the differences between language and a number of project decisions discussed below, the present results are not suitable for comparison with the work done by the actual task participants, and it should be regarded simply as an ongoing effort to generate and evaluate Portuguese descriptions using similar standards.

## 2 System Description

Our work is a simple application of n-gram statistics to surface realisation. Two independent annotators started by producing individual lists of the most likely phrases that could possibly be associated with every attribute–value pair in the corpus. Since at this initial stage we are only considering 1-to-n relations (i.e., each phrase is the realisation of exactly one attribute-value pair) the mapping annotation was straightforward. More complex (m-to-n) cases – those in which two or more properties may combine to form a single text unit (e.g., the properties of being human, young and male may be realised simply as “a young man” or even as “a boy”) – will be discussed elsewhere.

Given a TUNA attribute set as an input, we compute all (unordered) sets of phrases that correspond to a possible description, including gender variations. Next, we compute all possible permutations of each phrase set that matched a pre-defined description template suitable to Portuguese phrase order, once again with gender variation. As a result, even a simple attribute set as in “the large red table” would have at least eight possible realisations in Portuguese, although only a few can be considered well-formed and likely to be uttered for the purpose of identification. The final task of selecting the most likely output string is left to a simple bigram language model built from a 40-million words corpus of Brazilian Portuguese newspaper articles.

## 3 Preliminary Evaluation

We produced a surface realisation form for each of the 80 instances of Portuguese descriptions in the REG-2008 development data. Overall, 32 instances (40%) of descriptions were incorrectly generated. The major source of errors was the lack of complete gender agreement, since our simple bigram-based model cannot handle long-distance dependencies appropriately, as in “o sofá grande vermelha”, in which the gender agreement between “sofá” (masculine) and “vermelha” (feminine) could not be established. We believe that this could be easily fixed had we used a more expressive language model instead.

Two independent annotators built a Portuguese reference set by manually translating each of the 80 descriptions in the development data set and taking into account the possible phrase realisations defined earlier. More specifically, we produced a ‘normalized’ reference set, removing much of the noise that naturally occurs in the raw data. This included a number of likely errors (e.g., “red chair in center red”), meta-attributes (e.g., “first picture on third row”), illegal attributes (e.g.., “the grey desk with drawers”), differences in specificity (e.g., “shown from the side” as a less specific alternative to both “facing left” and “facing right” values) and synonymy (e.g., “facing the viewer” as an alternative to “facing forward”). Moreover, given that definiteness cannot be worked out from the attribute set alone, all indefinite descriptions were changed to definite.

Regarding the usefulness of this modified reference set, there are a number of due observations: firstly, given the differences between languages, our reference data set is not to be regarded as a resource for investigating language use as the original TUNA data set is intended to be, but rather as a standard of acceptable performance for a practical Portuguese NLG system. Moreover, since the translated descriptions were not produced in real situations of reference, we are aware that our results are  not directly comparable to, e.g., the work carried out in the REG-2008 challenge for evaluating English descriptions, and that would remain the case even without normalization.

On the other hand, although the result of both translation and normalization tasks is a somewhat simplified set of Portuguese descriptions, this is not to say that these descriptions are tailored to match those that we intend to generate. In fact, one of the goals in the normalization task was to retain the most appropriate instances of reference, which included a large number of cases that we are not presently able to produce, e.g., those combining the *x-dimension* and *y-dimension* attributes in single references to corners, as in “in the upper right corner”. Figure 1 summarizes our findings for the 80 instances of descriptions in the Furniture domain.

| | Furniture |
|---|---|
| String Accuracy | 0.26 |
| String-edit dist. | 3.26 |

Figure 1. Portuguese descriptions (Furniture domain)

## 4 Final Remarks

One striking difference between system descriptions and the reference set was the word order of Portuguese adjectives. To our surprise, it is not clear in which order attributes such as *colour* and *size* should be combined in Portuguese definite descriptions. For example, “*a large red table*” could be realised either as *type* + *colour* + *size* (e.g., “a mesa vermelha, grande” ) or as *type* + *size* + *colour* (e.g., “a mesa grande, vermelha”). As both alternatives seem equally acceptable, the choice may depend on which property contrasts each of the distractors in the situation of reference. Whilst the present ambiguity reveals a weakness in our artificially-built reference set, it may also 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 standard. We believe that further investigation on this issue is still required

## Acknowledgments

This work has been partially supported by CNPq-Brazil (484015/2007-9), FAPESP (2006/03941-7) and Ensinar com Pesquisa / USP.

## References

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

van Deemter, K.; I. van der Sluis and A. Gatt (2006) Building a semantically transparent corpus for the generation of referring expressions. 4th International Conference on Natural Language Generation.