

# Referring via Document Parts

Ivandr  Paraboni<sup>1</sup> and Kees van Deemter<sup>2</sup>

<sup>1</sup> Instituto de Ci ncias Matem ticas e de Computa  o - ICMC  
Universidade de S o Paulo - USP  
Av Trabalhador S o-Carlense, 400, 13560-970 - S o Carlos SP, Brazil  
ivandre@icmc.usp.br

<sup>2</sup> Department of Computing Science, King's College, University of Aberdeen  
Aberdeen AB24 3UE, Scotland, UK  
kvdeemte@csd.abdn.ac.uk

**Abstract.** Documents in a wide range of genres often contain references to their own sections, pictures etc. We call such referring expressions instances of *Document Deixis*. The present work focuses on the generation of Document Deixis in the context of a particular kind of natural language generation system in which these descriptions are not specified as part of the input, i.e., when it is up to the system to decide whether a reference is called for and, if so, which document entity it should refer to. We ask under what circumstances it is advantageous to describe domain objects in terms of the document parts where they are mentioned (as in “the insulin described in section 2”). We report on an experiment suggesting that such indirect descriptions are preferred by human readers whenever they cause the generated descriptions to be shorter than they would otherwise be.

## 1 Introduction

Document parts such as sections, subsections, pictures, paragraphs etc may be referred to for various purposes, for example to point to additional information on the current topic of the text, e.g., “see also section 7”. References to document parts will often be *deictic*, in the sense that the realisation of the expression depends on the place in the document where the referring expression is uttered (e.g., “this section” versus “section 1.1.”). Accordingly, we will call the references to parts of the *same* document instances of Document Deixis (DDX).

We are interested in a particular kind of DDX, which we have previously called *object-level* instances of Document Deixis [9]. These are usually part of a larger expression which refers to a domain entity. The entity in question may be concrete (e.g., the medicines in Example 1) or abstract (e.g., the advice in Example 2). In the corpora that we investigated – patient information leaflets [1] - references to abstract entities or sets of them are far more common.