

# Preface

This book presents several contributions oriented towards the common goal of bridging the gap between Sikkel's theory of *parsing schemata* and the practical aspects of natural language parser development.

On the practical side, a compiler is presented that can be used to automatically generate an efficient implementation of a parser from its formal description in the form of a parsing schema. This system is then used to obtain implementations of several well-known parsers for context-free grammars and tree-adjoining grammars, test them with practical natural language grammars and then conduct an analysis to determine their empirical performance.

On the theoretical side, two extensions of parsing schemata are introduced, enabling the formalism to describe two kinds of parsers that are useful in practical natural language processing applications, but were not previously supported by this theory.

The first extension is for *error-repair parsers*, which are algorithms able to robustly parse ungrammatical sentences. Apart from the extension itself, a transformation is presented that can be used to automatically obtain error-repair parsers from standard parsers.

The second extension defines a variant of parsing schemata for *dependency parsers*, which are algorithms that represent the structure of sentences as a set of links between their words. This formalism is used to compare and relate several well-known projective and non-projective dependency parsers, as well as to solve the problem of efficiently parsing *mildly non-projective dependency structures* by defining novel algorithms for several sets of these structures.

Put together, the results in this book provide the parser developer with a common formal framework that can be used to design, analyse and

compare different kinds of parsing algorithms, including error-repair and dependency-based parsers; as well as practical tools to automatically obtain efficient implementations of these parsers directly from their formal representation.

This book is based on my Ph.D. thesis, completed in 2009 at the University of Corunna. However, extensive changes, additions and improvements have been made to make it useful to a wider range of readers. Many of the contributions presented here were originally published as journal articles or papers in conference proceedings, as can be seen in the references. Thus, first and foremost, I would like to acknowledge and thank my co-authors in these publications (Miguel A. Alonso, John Carroll, Jesús Vilares, Manuel Vilares and David Weir) for their valuable contributions and collaboration. I am also grateful to Víctor J. Díaz, Joakim Nivre, Giorgio Satta and Leo Wanner for the helpful comments and suggestions that they provided as reviewers of the thesis; as well as to the anonymous referees that reviewed the earlier draft of this book and the publications from which it draws. I am also indebted to Klaas Sikkel, not only for his helpful comments on the draft, but also for developing the theory of parsing schemata which is the starting point and foundation of this book. The research reported in the book has been supported in part by grants from the Spanish *Ministerio de Educación y Ciencia* and *FEDER*<sup>1</sup> and *Xunta de Galicia*.<sup>2</sup>

The book can be read by researchers, graduate students and post-graduates interested in computational linguistics and natural language engineering. The reader is assumed to have an elementary background in mathematics and computing, as provided by undergraduate computer science programs. Previous knowledge of parsing or computational linguistics is helpful but by no means required, since all the concepts needed are presented either in the introductory chapter or in the parts of the book where they are relevant. Thus, the book may be used as teaching material for a course on natural language parsing, by starting with the first part and then choosing among the other parts depending on the desired goals, since they can be read independently of each other.

*Carlos Gómez-Rodríguez*

---

<sup>1</sup>Projects TIN2004-07246-C03-01, HUM2007-66607-C04-03 and Progr. de Becas FPU.

<sup>2</sup>Grants PGIDIT05PXIC10501PN, PGIDIT07SIN005206PR, Estadías do programa de RRHH da Dirección Xeral de I+D+i, Redes Galegas de Procesamento da Linguaxe e Recuperación de Información e de Lingüística de Corpus.