

Technologies for the Processing and Retrieval of Semi-Structured Documents

Experience from the CADIAL Project

Edited by

Marko Tadić

Bojana Dalbelo Bašić

Marie-Francine Moens

Croatian Language Technologies Society
Zagreb 2009

The Computer-Aided Document Indexing for Accessing Legislation
(CADIAL) project was jointly supported by:



Flemish Government
Department International Flanders



Croatian Government
Ministry of Science, Education and Sports

© 2009 Croatian Language Technologies Society
and several chapters also by respective publishers as marked in notes.

A CIP catalogue record for this book is available from
the National and University Library in Zagreb under number 713315
ISBN 978-953-55375-1-9

Contents

Preface

Ralf Steinberger

vii

Introduction

Chapter 1

Computer Aided Document Indexing Accessing Legislation:
A Joint Venture between Flanders and Croatia

Bojana Dalbelo Bašić, Marie-Francine Moens, Marko Tadić

3

Chapter 2

HIDRA's Motivation and Role in the CADIAL Project

Neda Erceg, Maja Cvitaš

15

Language Technologies for Information Retrieval

Chapter 3

Lexicon-Based Morphological Normalisation and its Application
to Croatian Language

Jan Šnajder, Bojana Dalbelo Bašić, Marko Tadić

23

Chapter 4

Higher-Order Functional Representation of Croatian Inflectional
Morphology

Jan Šnajder, Bojana Dalbelo Bašić

81

Chapter 5

Implementation of the Croatian NERC System

Božo Bekavac, Marko Tadić

99

Chapter 6

A Generic Method for Multi Word Extraction from Wikipedia

Božo Bekavac, Marko Tadić

115

Chapter 7

Evolving New Lexical Association Measures Using Genetic
Programming

Jan Šnajder, Bojana Dalbelo Bašić, Saša Petrović, Ivan Sikirić

125

<i>Chapter 8</i>		
Evaluating Full Lemmatization of Croatian Texts		
Željko Agić, Marko Tadić, Zdravko Dovedan		133
<i>Chapter 9</i>		
TermeX: A Tool for Collocation Extraction		
Davor Delač, Zoran Krleža, Jan Šnajder, Bojana Dalbelo Bašić, Frane Šarić		145
Knowledge Technologies for Information Retrieval		
<i>Chapter 10</i>		
Feature Selection for Document Categorization		
Erik Boiy, Marie-Francine Moens		159
<i>Chapter 11</i>		
Comparing Document Classification Schemes Using K-Means Clustering		
Artur Šilić, Marie-Francine Moens, Lovro Žmak, Bojana Dalbelo Bašić		177
<i>Chapter 12</i>		
Building a Search Engine Model with Morphological Normalization Support		
Jure Mijić, Bojana Dalbelo Bašić, Jan Šnajder		191
<i>Chapter 13</i>		
TMT: Object-Oriented Text Classification Library		
Artur Šilić, Frane Šarić, Bojana Dalbelo Bašić, Jan Šnajder		201
<i>Chapter 14</i>		
CADIAL Search Engine at INEX		
Jure Mijić, Marie-Francine Moens, Bojana Dalbelo Bašić		213
<i>Chapter 15</i>		
Accessing Information in Semi-Structured Documents: Research Avenues		
Marie-Francine Moens		225