Preface

Most, if not all public authorities such as parliaments and public administrations classify and catalogue their legal, administrative and other documents in order to allow their users efficient search and retrieval of relevant texts. Unlike on the internet, where web documents are full-text indexed so that users can find all web pages containing certain words, public authorities additionally index using a given list of concepts from a thesaurus. The advantage is that users can search for a concept like 'EQUALITY OF MEN AND WOMEN' and will find the relevant legislation even if the exact search string is not present in the text. When a multilingual thesaurus is used, users can even retrieve relevant documents in other languages while entering thesaurus search terms in their own language. This book addresses conceptual indexing and related language technology applications.

A major standard to index legislative and administrative documents is the Eurovoc thesaurus, used by the European Parliament, the European Commission and many national public authorities in the European Union (EU) and beyond. Eurovoc has been in use since 1984 and it is being maintained and updated continuously. Eurovoc distinguishes over six thousand categories, so called descriptors. Human indexers typically assign between three and ten of the most relevant descriptors to each document. To date, Eurovoc exists in altogether 30 language translations (22 official languages of the European Union plus several others, including Croatian and Russian). Using this same multilingual thesaurus for document classification allows information search and exchange across different organisations, countries and even languages. The fact that professionals and civilians are able to access public documents is an important feature for a transparent democracy. Being able to search and access public documents in other EU countries brings the countries closer together and favours a quicker integration of the many EU Member States.

While the benefits of conceptual indexing are obvious, there is significant human and financial cost involved in providing it. Professional documentalists index an average of about thirty documents per day, while their organisations may need hundreds of documents to be classified. Human indexing is time-consuming and thus expensive. Due to the tediousness of the task, staff turnover is high and indexing quality is irregular. Furthermore, concepts change rather quickly so that older documents should ideally be re-indexed every few years. The usage of automatic indexing software would thus be an obvious solution. For this reason, in the mid-1990s, the European Parliament instigated two studies to evaluate automatic procedures to Eurovoc-classify documents. However, the indexing quality of the automatic software was too low so that they had to abandon the idea. Starting in 2001, the European Commission's Joint Research Centre tackled the task again, originally not with the objective to support documentalists in their work, but to automatically link related web documents with each other across ten languages. In the meantime, this software has been trained for twenty-two European languages and it runs daily and successfully as part of a multilingual news analysis system. However, the indexing quality is still too low for the purposes of professional parliamentary use. Only the Spanish Congress of Deputies in Madrid actually uses this software for the daily indexing of their documents, solving the accuracy issue by using the system interactively.

The CADIAL project (Computer-Aided Document Indexing for Accessing Legislation) had the aim of tackling the issue of automatic Eurovoc indexing yet again, focusing this time mainly on three specific issues: (1) dealing with a morphologically complex language such as Croatian, an issue not considered in the previous approaches; (2) considering – since the beginning – the end users' needs for a semi-automatic, i.e., interactive, computer-aided document indexing system, by working closely with the Croatian Information Documentation Referral Agency HIDRA; and (3) exploring a variety of technical variations regarding the document representation and the classification algorithms. These latter experiments are important because – while text classification as such is a well-known task – Eurovoc indexing is particularly difficult due to the high number of descriptors (over six thousand) and their highly imbalanced distribution, as well as the text type differences in the targeted document collection.

This book summarises all aspects of the efforts of the CADIAL project for automatic document indexing. The first part (Chapters 1 and 2) gives a detailed overview of the contents of this book, provides a historical overview of previous initiatives, and presents the task from a user's perspective. The second part (Chapters 3 to 9) describes document representation issues, including automatic procedures to morphologically normalise words in highly inflected languages, the recognition of named entities such as persons, organisations and locations in texts, as well as the identification and usage of multi-word terms and collocations. The third part includes discussions on document classification algorithms and search engine features to map user requests with the automatic assignment results. Chapter 15 rounds up the discussion with an outlook on further research avenues.

The book will thus be an interesting read not only for language technology specialists, but also for end users interested in the domain of automatic document indexing, library sciences, information retrieval, and more.

Ralf Steinberger
European Commission
Joint Research Centre
Ispra, Italy