# Preface

Ambiguity and variability are two basic and pervasive phenomena characterizing lexical semantics. In this book we introduce a computational model for lexical semantics based on Semantic Domains. This concept is inspired by the "Theory of Semantic Fields", proposed in structural linguistics to explain lexical semantics. The main property of Semantic Domains is lexical coherence, i.e. the property of domain-related words to co-occur in texts. This allows us to define automatic acquisition procedures for Domain Models from corpora, and the acquired models provide a shallow representation for lexical ambiguity and variability. Domain Models have been used to define a similarity metric among texts and terms in the Domain Space, where second-order relations are reflected. Topic similarity estimation is at the basis of text comprehension, allowing us to define a very general domain-driven methodology. The basic argument we put forward to support our approach is that the information provided by the Domain Models can be profitably used to boost the performances of supervised Natural Language Processing systems for many tasks. In fact, Semantic Domains allows us to extract domain features for texts, terms and concepts. The obtained indexing, adopted by the Domain Kernel to estimate topic similarity, preserves the original information while reducing the dimensionality of the feature space. The Domain Kernel is used to define a semi-supervised learning algorithm for Text Categorization that achieves state-of-the-art results while decreasing by one order the quantity of labeled texts required for learning. The property of the Domain Space to represent together terms and texts allows us to define an Intensional Learning schema for Text Categorization, in which categories are described by means of discriminative words instead of labeled examples, achieving performances close to human agreement. Then we investigate the role of domain information in Word Sense Disambiguation, developing both unsupervised and supervised approaches that strongly rely on the notion of Semantic Domain. The former is based on the lexical resource WordNet Domains and the latter exploits both sense tagged and unlabeled data to model the relevant domain distinctions among word senses. The proposed supervised approach improves the

state-of-the-art performance in many tasks for different languages, while reducing appreciably the amount of sense tagged data required for learning. Finally, we present a lexical acquisition procedure to obtain Multilingual Domain Models from comparable corpora. We exploit such models to approach a Cross-language Text Categorization task, achieving very promising results.

We would first of all acknowledge the effort of other people involved in the eight years' long daily work required to produce the experimental results reported in this monograph, and in particular Claudio Giuliano, who performed most of the experimental work for the WSD experiments, allowing us to achieve very accurate results in competitions due to his patience and skills; to Bernardo Magnini, who first proposed the concept of Semantic Domain, opening the direction we have followed during our research path and supporting it with financial contributions from his projects; and to Ido Dagan, who greatly contributed to the intensional learning framework defining the experimental settings and clarifying the statistical properties of the GM algorithm.

Special thanks are devoted to Oliviero Stock, for his daily encouragement and for the appreciation he has shown for our work; to Walter Daelemans, who demonstrated a real interest in the epistemological aspects of this work from the early stages; to Maurizio Matteuzzi, whose contribution was crucial to interpret the theoretical background of this work related to philosophy of language; to Roberto Basili who immediately understood the potential of Semantic Domains and creatively applied our framework for technology transfer, contributing to highlighting limitations and potentialities; and to Aldo Gangemi, who more recently helped us in clarifying the relationship of this work with formal semantics and knowledge representation.

Last, but not least, we would like to thank our families and parents for having understood with patience our crazy lives, and our friends for having spent their nights in esoteric and sympathetic discussions.

Trento, *Alfio Gliozzo*
September 2008 *Carlo Strapparava*