Preface

Grammatical inference is one part of theoretical computer science that addresses the problem of how computers can learn from experience. In this way, it is like the well-known field of machine learning. Perhaps what most distinguishes grammatical inference from standard machine learning approaches is its focus on learning the structure which underlies the concept to be learned, i.e., in identifying the *nature* of the *target* concept.

A common ingredient that one can find in all the grammatical inference works is that the target concepts are formal objects like sets of strings or sets of trees, which can be represented either by accepting devices (for example, a finite automata, a neural network, etc.) or by generating devices (mainly, formal grammars which can be inserted into a formal framework such as the *Chomsky hierarchy*). Hence, it is common for the results and algorithms developed in the area of grammatical inference to have implications in other research areas such as computability and complexity theory, formal languages and applications, artificial intelligence, etc. Similarly, such areas frequently guide research in grammatical inference. So, grammatical inference and other research areas are continuously influencing each other.

Grammatical inference has been developed largely in the last 30 years. We can establish E.M. Gold's work *Language Identification in the Limit* (1967) as the seminal work where all the basic problems, concepts, and rules of the game were established for grammatical inference. Since then, the grammatical inference research area has been continuously in progress, especially since the landmark work in the 1980s by Dana Angluin. So, it is a well-developed area with a long tradition.

There is a well-established community of researchers who have a presence in the scientific scene through different publications and conferences. The main conference on grammatical inference is international in scale, and has been celebrated every 2 years since 1993 (the ICGI conference series). Its proceedings were published as part of the Springer LNCS or LNAI series until 2010, and since 2012 they have been published as part of the open-access JMLR Workshop and Conference

Proceedings series.[1] These conferences are directed by an international steering committee (of which we are current members). Additionally, the website http://www.grammarlearning.org/ provides online information about conferences, competitions, software, and other resources relating to the field of grammatical inference. Lastly, the 2010 book *Grammatical Inference* by Colin de la Higuera, published by Cambridge University Press, collects in one place the main body of results of this field and presents them from the ground up in a uniform fashion.

This book provides advanced treatments of topics in grammatical inference. In this way, this book complements de la Higuera's book, by addressing topics that are either unmentioned or only introduced there.

The topics included in this book are largely drawn from tutorials that were presented at the ICGI 2010 and 2012 conferences. They were selected for the following reasons: (1) the topic has reached a certain level of scientific maturity, so a reasonable number of (positive and negative) results, algorithms, and conclusions have been obtained; and (2) the topic is of fundamental interest to grammatical inference, so it attracts a significant number of researchers to study the many facets of the problems that it exhibits.

The first three chapters of the book deal with issues regarding the theoretical learning framework. So, John Case's chapter, Gold-Style Learning Theory, discusses different learning paradigms, relationships among them, and learning power associated with them more generally. Rémi Eyraud, Jeffrey Heinz, and Ryo Yoshinaka's chapter, *Efficiency in the Identification in the Limit Learning Paradigm,* pays special attention to the complexity issues associated with *identification in the limit* learning criteria. Then, a chapter by Colin de la Higuera, *Learning Grammars and Automata with Queries*, focuses on the main results of learning by changing the information source to what is called *active learning*, wherein the learner can ask an oracle for information about the target. The next part of the book focuses on the main classes of formal languages according to Chomsky's hierarchy: the regular languages and the context-free languages. With respect to the regular languages, two chapters deal with finite-state automata. First, the chapter by Damián López and Pedro Garca, *On the Inference of Finite State Automata from Positive and Negative Data,* shows the main aspects of learning regular languages from examples and counterexamples. The chapter by Jorge Castro and Ricard Gavaldà, *Learning Probability Distributions Generated by Finite-State Machines,* approaches the learning of stochastic regular languages in a probabilistic manner, with a special focus on spectral learning. The chapter *Distributional Learning of Context-Free and Multiple Context-Free Grammars,* by Alexander Clark and Ryo Yoshinaka, focuses on an algebraic approach to learning some subclasses of the context-sensitive languages, which include significant classes of context-free languages. The next chapter, by Johanna Björklund and Henning Fernau, *Learning Tree Languages*, largely deals with the learning of regular sets of tree languages. The relation between tree languages and context-free

---

[1]http://jmlr.csail.mit.edu/proceedings/.

languages of strings was established as an alternative approach to learn in what is named *learning from structural data*. Hence, most of the results of this chapter are relevant to learning context-free languages. Finally, the chapter by François Coste, *Learning the Language of Biological Sequences,* shows an area of application that has recently been approached by grammatical inference: the processing of biosequences.

One decision we made early in regards to this book was to give the authors a high level of autonomy in preparing their chapters. There are advantages and disadvantages to this approach. One disadvantage is that some overlap inevitably exists between the chapters. We have cross-referenced other chapters where appropriate. Another disadvantage is that the notation used in each chapter differs. However, we believe the advantages outweigh these disadvantages. Each chapter in this book stands on its own, and includes the concepts and references necessary to help the understanding of the results by the reader. Thus it is not necessary for the reader to approach the contents of this book in order. Additionally, as a consequence of this approach, this book is oriented to an audience with basic knowledge of mathematics, computer science, and formal language theory. It could be (under)graduate students, or computer scientists, linguistics researchers, cognitive psychologists, or other readers interested in the nature of learning and its relation to computer science, artificial intelligence, and a significant number of related areas.

The editing of this book has been a long process where we have been helped by many different persons. We would like to thank them all for the support that they have provided during this process. First, we would like to thank all the contributors and authors of this book for the patience that they have exhibited during all stages of the book's production, especially during the reviewing process. We specially thank Colin de la Higuera for his support and for giving us the idea of editing this book. We thank all the people from Springer for their support, specially Ronan Nugent for all his understanding about the delays and problems that we encountered during the preparation of this book. Last but not least, we give thanks to our family members and friends for all their support. This book has been a good conversation topic during all this time, although they avoided asking much about it on some occasions. To all of them, thank you very much.

Newark, DE, USA   Jeffrey Heinz
Valencia, Spain   José M. Sempere
December 2014