Preface

Motivation and Subject

Language processors have become an inseparable part of our daily life. For instance, all the sophisticated modern means of communication, such as Internet with its numerous information processing tools, are based upon them to some extent, and indisputably, literally billions of people use these means on a daily basis. It thus comes as no surprise that the scientific development and study of languages and their processors fulfill a more important role today than ever before. Naturally, we expect that this study produces concepts and results that are as reliable as possible. As a result, we tend to base this study upon mathematics as a systematized body of unshakable knowledge obtained by exact and infallible reasoning. In this respect, we pay our principal attention to *formal language theory* as a branch of mathematics that formalizes languages and devices that define them strictly rigorously.

This theory defines languages mathematically as sets of sequences consisting of symbols. This definition encompasses almost all languages as they are commonly understood. Indeed, natural languages, such as English, are included in this definition. Of course, all artificial languages introduced by various scientific disciplines can be viewed as formal languages as well; perhaps most illustratively, every programming language represents a formal language in terms of this definition. Consequently, formal language theory is important to all the scientific areas that make use of these languages to a certain extent.

The strictly mathematical approach to languages necessitates introducing *formal language models* that define them, and formal language theory has introduced a great variety of them over its history. Most of them are based upon rules by which they repeatedly rewrite sequences of symbols, called strings. Despite their diversity, they can be classified into two basic categories—generative and recognition language models. Generative models, better known as *grammars*, define strings of their language and so their rewriting process generates them from a special start symbol. On the other hand, recognition models, better known as *automata*,

define strings of their language by rewriting process that starts from these strings and ends in a special set of strings, usually called *final configurations*.

Like any branch of mathematics, formal language theory has defined its language models generally. Unfortunately, from a practical viewpoint, this generality actually means that the models work in a completely non-deterministic way, and as such, they are hardly implementable and, therefore, applicable in practice. Being fully aware of this pragmatic difficulty, formal language theory has introduced fully deterministic versions of these models; sadly, their application-oriented perspectives are also doubtful. First and foremost, in an ever-changing environment in which real language processors work, it is utterly naive, if not absurd, that these deterministic versions might adequately reflect and simulate real language processors applied in such pragmatically oriented areas as various engineering techniques for language analysis. Second, in many case, this determinism decreases the power of their general counterparts—another highly undesirable feature of this strict determinism.

Considering all these difficulties, formal language theory has introduced yet another version of language models, generally referred to as *regulated language models*, which formalize real language processors perhaps most adequately. In essence, these models are based upon their general versions extended by an additional mathematical mechanism that prescribes the use of rules during the generation of their languages. From a practical viewpoint, an important advantage of these models consists in controlling their language-defining process and, therefore, operating in a more deterministic way than general models, which perform their derivations in a quite unregulated way. Perhaps even more significantly, the regulated versions of language models are stronger than their unregulated versions. Considering these advantages, it comes as no surprise that formal language theory has paid an incredibly high attention to *regulated grammars and automata*, which represent the principal subject of the present book.

Purpose

Over the past quarter century, literally hundreds of studies were written about regulated grammars, and their investigation represents an exciting trend within formal language theory. Although this investigation has introduced a number of new regulated grammatical concepts and achieved many remarkable results, all these concepts and results are scattered in various conference and journal papers. The principal *theoretical purpose* of the present book is to select crucially important concepts of this kind and summarize key results about them in a compact, systematic, and uniform way.

From a more practical viewpoint, as already stated, the developers of current and future language processing technologies need a systematized body of mathematically precise knowledge upon which they can rely and build up their methods and techniques. The *practical purpose* of this book is to provide them with this knowledge.

Focus

The material concerning regulated grammars and automata is so huge that it is literally impossible to cover it completely. Considering the purpose of this book, we restrict our attention to four crucially important topics concerning these grammars and automata—their power, properties, reduction, and convertibility.

As obvious, the *power* of the regulated language models under consideration represents perhaps the most important information about them. Indeed, we always want to know the family of languages that these models define.

A special attention is paid to algorithms that arrange regulated grammars and automata and so they satisfy some prescribed *properties* while the generated languages remain unchanged because many language processors strictly require their satisfaction in practice. From a theoretical viewpoint, these properties frequently simplify proofs demonstrating results about these grammars and automata.

The *reduction* of regulated grammars and automata also represents an important investigation area of this book because their reduced versions define languages in a succinct and easy-to-follow way. As obvious, this reduction simplifies the development of language processing technologies, which then work economically and effectively.

Of course, the same languages can be defined by different language models. We obviously tend to define them by the most appropriate models under given circumstances. Therefore, whenever discussing different types of equally powerful language models, we also study their mutual *convertibility*. More specifically, given a language model of one type, we explain how to convert it to a language model of another equally powerful type and so both the original model and the model produced by this conversion define the same language.

We prove most of the results concerning the topics mentioned above *effectively*. That is, within proofs demonstrating them, we give algorithms that describe how to achieve these results. For instance, we often present conversions between equally powerful models as algorithms, whose correctness is then rigorously verified. In this way, apart from their theoretical value, we actually demonstrate how to implement them.

Organization

The text is divided into nine parts, each of which consists of several chapters. Every part starts with an abstract that summarizes its chapters. Altogether, the book contains twenty-two chapters.

Part I, consisting of Chaps. 1 through 3, gives an introduction to this monograph in order to express all its discussion clearly and, in addition, make it completely self-contained. It places all the coverage of the book into scientific context and reviews important mathematical concepts with a focus on formal language theory.

Part II, consisting of Chaps. 4 and 5, gives the fundamentals of regulated grammars. It distinguishes between context-based regulated grammars and rule-based regulated grammars. First, it gives an extensive and thorough coverage of regulated grammars that generate languages under various context-related restrictions. Then, it studies grammatical regulation underlain by restrictions placed on the use of rules.

Part III, consisting of Chaps. 6 through 9, covers special topics concerning grammatical regulation. First, it studies special cases of context-based regulated grammars. Then, it discusses problems concerning the erasure of symbols in strings generated by regulated grammars. Finally, this part presents an algebraic way of grammatical regulation.

Part IV, consisting of Chaps. 10 through 12, studies parallel versions of regulated grammars. First, it studies generalized parallel versions of context-free grammars, generally referred to as regulated ET0L grammars. Then, it studies how to perform the parallel generation of languages in a uniform way. Finally, it studies algebraically regulated parallel grammars.

Part V, consisting of Chaps. 13 and 14, studies sets of mutually communicating grammars working under regulating restrictions. First, it studies their regulation based upon a simultaneous generation of several strings composed together by some basic operation after the generation is completed. Then, it studies their regulated pure versions, which have only one type of symbols.

Part VI, consisting of Chaps. 15 and 16, presents the fundamentals of regulated automata. First, it studies self-regulating automata. Then, it covers the essentials concerning automata regulated by control languages.

Part VII, consisting of Chaps. 17 and 18, studies modified versions of classical automata closely related to regulated automata—namely, jumping finite automata and deep pushdown automata.

Part VIII, consisting of Chaps. 19 and 20, demonstrates applications of regulated language models. It narrows its attention to regulated grammars rather than automata. First, it describes these applications and their perspectives from a rather general viewpoint. Then, it adds several case studies to show quite specific real-world applications concerning computational linguistics, molecular biology, and compiler writing.

Part IX, consisting of Chaps. 21 and 22, closes the entire book by adding several remarks concerning its coverage. First, it sketches the entire development of regulated grammars and automata. Then, it points out many new investigation trends and long-time open problems. Finally, it briefly summarizes all the material covered in the text.

Approach

This book represents a theoretically oriented treatment of regulated grammars and automata. We introduce all formalisms concerning these grammars with enough rigor to make all results quite clear and valid. Every complicated mathematical

passage is preceded by its intuitive explanation so that even the most complex parts of the book are easy to grasp. As most proofs of the achieved results contain many transformations of regulated grammars and automata, the present book also maintains an emphasis on algorithmic approach to regulated grammars and automata under discussion and, thereby, their use in practice. Several worked-out examples illustrate the theoretical notions and their applications.

Use

Primarily, this book is useful to all researchers, ranging from mathematicians through computer scientists up to linguists, who deal with language processors based upon regulated grammars or automata.

Secondarily, the entire book can be used as a text for a two-term course in regulated grammars and automata at a graduate level. The text allows the flexibility needed to select some of the discussed topics and, thereby, use it for a one-term course on this subject.

Tertiarily and finally, serious undergraduate students may find this book useful as an accompanying text for a course that deals with formal languages and their models.

WWW Support

Further backup materials, such as lectures about selected topics covered in the book, are available at

http://www.fit.vutbr.cz/~meduna/books/rga

Brno, the Czech Republic Alexander Meduna
Brno, the Czech Republic Petr Zemek

http://www.springer.com/978-1-4939-0368-9

Regulated Grammars and Automata
Meduna, A.; Zemek, P.
2014, XX, 694 p. 12 illus., Hardcover
ISBN: 978-1-4939-0368-9