Contents

Introduction

This book is a quite extended version of lectures basically dedicated to combinatorics on words and numeration systems that I am giving at the University of Liège. The course is usually (but not necessarily) followed by students interested in discrete mathematics or theoretical computer science. The chosen level of abstraction should allow undergraduate students to study the exposed topics.

What this book is or is not about

In the long process of writing this book, I have expanded my initial notes with many examples and many extra concepts to have a somehow consistent overview of the field. Nevertheless, this book is *not* intended to serve as an encyclopedic reference.

I have picked some of my favorite topics in the area and I have also decided to shorten the presentation of some items (not because there are less interesting but choices have to be made to keep this book reasonably short). Indeed, the book most probably reflects what I do prefer: I'm always more interested in the *combinatorics* and the underlying *discrete structures* arising from a problem.

When preparing this book, I have chosen to present a fairly large variety of basic notions and important tools and results. Sometimes, I only give an overview of a subject and *proofs are therefore omitted*. For the reader wanting to study further a specific topic, many pointers to the relevant bibliography are given and each chapter ends with notes and comments. Indeed, the main goal of this book is to give *a quick access to actual research topics at the intersection between automata and formal language theory, number theory and combinatorics on words*.

A few words about what you will find

The notion of a word, i.e., a (finite or infinite) sequence of symbols belonging to a finite set, is central all along this book. It has connections with many branches of mathematics and computer science: number theory, combinatorics, formal language theory, mathematical logic, symbolic dynamics, coding theory, computational complexity, discrete geometry, stringology, etc.

Combinatorics on words. One can be interested in the combinatorial properties of finite or infinite sequences of symbols over a finite alphabet: what are the possible arrangements, how many such configurations can be achieved,... As a trivial example, over a binary alphabet any word of length at least 4 contains a repeated factor of the kind uu (try to prove it!). One can therefore look at patterns that are unavoidable in sufficiently long sequences or count the number of patterns or configurations that may appear in a particular context. These are some of the general questions that will be considered in the first three chapters of this book. In particular, we shall concentrate on infinite words that can be obtained by a simple procedure consisting in the iteration of a morphism over a free monoid. We shall mostly deal with a large class of self-similar words: the so-called morphic words and, in particular, with automatic sequences that are generated by a constant-length morphism.

Formal language theory. A language is merely a set of words. In this book, we shall mostly encounter languages of finite words. One exception is a short incursion into symbolic dynamical systems with the language of the β-expansions of the real numbers in the interval $[0, 1)$. The Chomsky's hierarchy introduced in the theory of formal languages provides a classification depending on the machine needed to recognize an infinite language of finite words. From a computational perspective, the simplest languages are the regular languages. They are accepted (or recognized) by finite automata, and described by regular expressions. Chapter 4 is a short chapter presenting the main properties of these languages. We shall constantly see connections existing between regular languages, automatic sequences and numeration systems. For instance, we associate quite often a finite automaton with a morphism.

Number theory. A finite word can also be used to represent an integer in a given numeration system (e.g., integer base expansions and many other non-standard systems are discussed in depth in several chapters of this book). To quote A. Fraenkel: "There are many ways of representing an integer uniquely!" [FRA 85]. Similarly, an infinite word can represent a real number or the characteristic sequence of a set of integers. With that respect, a natural question is to study links existing between arithmetical properties of numbers (or sets of numbers) and syntactical properties of their expansions. Chapter 5 is dedicated to numeration systems with a particular emphasis put on words representing numbers. Indeed, the chosen numeration system has a strong influence on the syntactical properties of the corresponding representations. A cornerstone is the notion of recognizable set of numbers whose elements when represented within a given numeration system are recognized by a finite automaton.

Formal methods applied to infinite words and sets of numbers. In the last chapter of this book, we described a recent trend in combinatorics on words. Thanks to automata theory and Büchi's theorem, we shall see how formal methods enter the picture about decision problems or, automatic theorem-proving, relevant in combinatorics on words. If a property about some infinite words can be described by a well-written logical formula, then this property can be tested automatically. Such a procedure holds for a large class of infinite words generated by iterated morphisms (for automatic sequences and those stemming from Pisot numeration systems as presented in this book). The expressiveness of Presburger arithmetic (with an extra predicate) provides an interesting alternative to deal with a sufficiently large class of problems about infinite morphic words. One can imagine automated certificates for several families of combinatorial properties. But the price to pay is that we should have to deal with fairly large automata. It is a field of research where combinatorists and computer scientists can work together fruitfully: on the one hand, it is well-known that, in the worst-case, the obtained decision procedures can be super-exponential, but on the other hand, the considered problems about words seem to be of relatively small complexity.

How to read this book

The goal is that, after reading this book (or at least parts of this book), the reader should be able to fruitfully attend a conference or a seminar in the field. I hope that the many examples presented along the text will help the reader to get some feeling about the presented topics even though we are not going too far in the technical aspects of the proofs. Also, prerequisites are minimal. We shall not explore topics requiring measure theory or advanced linear algebra (we have avoided results related to Jordan normal form of matrices) or non-elementary number theory. Two sections are devoted to results in algebraic number theory and formal series. Sections 1.1.2 and 1.2.2 serve as references that the reader may consult when needed. Sections 6.1 and 6.2 give a self-contained presentation of the concepts of mathematical logic needed in this book. Those rigorous and technical sections should not discourage the reader to pursue his/her study. Most of the material can be accessed without much background.

My initial idea was to go quick to the point but it seems that the stories I wanted to tell were indeed quite longer than initially thought. I have to confess that writing this book was a quite unexpected adventure (I was perpetually trying to meet the deadlines and also dealing with my other duties at the University and at home).

There are several paths that the reader can follow through this book. Some are quite long, some are shorter.

– For a *basic introduction*, I propose to read parts of Chapter 1 (skipping the reference sections), Chapter 2 up to and including Section 2.4. If the reader has already some knowledge about automata, then one can conclude with Chapter 6 concentrating on results about integer base systems.

– For a one-semester course in *combinatorics on words*, I propose a reading of the first three chapters not sacrificing the rigorous presentation of Section 1.2.1.

– For a *numeration system* oriented reading, again organized over one semester: browse through the first chapter (with a careful reading of the examples related to numeration systems), then go to Section 2.3 and conclude with the last two chapters of the book.

– For a course oriented towards *interaction between automata, logic and numeration systems*, one can focus on Chapters 4 and 6.

About other sources treating similar subjects, an excellent companion for this book is definitely *Automatic Sequences: Theory, Applications, Generalizations* [ALL 03a] written by Allouche and Shallit. I do hope that the two books can be read independently and can benefit from each other. There is also a non-zero intersection with several chapters of the Lothaire's book *Algebraic Combinatorics on Words* (namely those about Sturmian words written by Berstel and Séébold and the one on numeration systems written by Frougny) [LOT 02]. Some chapters of the volume *Combinatorics, Automata and Number Theory* [BER 10] as well as [PYT 02] can also serve as a follow up for the present book. In particular, Cassaigne and Nicolas's chapter on factor complexity is a natural continuation for our Chapter 2. I should finish by mentioning two papers that were very influential in my work: [BRU 94] and [BRU 95]. With this book, I hope that the reader could learn as much material as the one found in these two papers.

Labels of bibliographic entries are based on the first three letters of the last name of the first author and then the year of publication. In the bibliography, entries are sorted in alphabetical order using these labels.

Acknowledgments

I would like to express my gratitude to Valérie Berthé for her constant and enthusiastic support, for the many projects we run together and finally, for her valuable comments on a draft of this book.

Several researchers have spent some precious time to read a first draft of this book, their careful reading, their feedback and expert comments were very useful and valuable: Anna Frid, Julien Leroy, Aline Parreau, Narad Rampersad, Eric Rowland, Aleksi Saarela and Jeffrey Shallit. They proposed many clever improvements of the text. I warmly thank them all. I would like to give a special thank to Véronique Bruyère for comments on the last chapter.

For their feedbacks, I also sincerely thank Jean-Paul Allouche, Émilie Charlier, Fabien Durand and Victor Marsault.

Even though he was not directly involved in the writing process of this book, the first half of the book has greatly benefited from the many discussions I had with Pavel Salimov when he was a post-doctoral fellow in Liège. Naturally, all the discussions and interactions I could have had along the years with students, colleagues and researchers worldwide had some great influence on me (but such a list would be too long) and I thank them all.

Michel Rigo.