
Preface to the Second Edition

As is fit, this second edition arose out of our readers' demands to read about new developments and our desire to write about them. Although parsing techniques is not a fast moving field, it does move. When the first edition went to press in 1990, there was only one tentative and fairly restrictive algorithm for linear-time substring parsing. Now there are several powerful ones, covering all deterministic languages; we describe them in Chapter 12. In 1990 Theorem 8.1 from a 1961 paper by Bar-Hillel, Perles, and Shamir lay gathering dust; in the last decade it has been used to create new algorithms, and to obtain insight into existing ones. We report on this in Chapter 13.

More and more non-Chomsky systems are used, especially in linguistics. None except two-level grammars had any prominence 20 years ago; we now describe six of them in Chapter 15. Non-canonical parsers were considered oddities for a very long time; now they are among the most powerful linear-time parsers we have; see Chapter 10.

Although still not very practical, marvelous algorithms for parallel parsing have been designed that shed new light on the principles; see Chapter 14. In 1990 a generalized LL parser was deemed impossible; now we describe two in Chapter 11.

Traditionally, and unsurprisingly, parsers have been used for parsing; more recently they are also being used for code generation, data compression and logic language implementation, as shown in Section 17.5. Enough. The reader can find more developments in many places in the book and in the Annotated Bibliography in Chapter 18.

Kees van Reeuwijk has — only half in jest — called our book “a reservation for endangered parsers”. We agree — partly; it is more than that — and we make no apologies. Several algorithms in this book have very limited or just no practical value. We have included them because we feel they embody interesting ideas and offer food for thought; they might also grow and acquire practical value. But we also include many algorithms that do have practical value but are sorely underused; describing them here might raise their status in the world.

Exercises and Problems

This book is not a textbook in the school sense of the word. Few universities have a course in Parsing Techniques, and, as stated in the Preface to the First Edition, readers will have very different motivations to use this book. We have therefore included hardly any questions or tasks that exercise the material contained within this book; readers can no doubt make up such tasks for themselves. The questions posed in the problem sections at the end of each chapter usually require the reader to step outside the bounds of the covered material. The problems have been divided into three not too well-defined classes:

- not marked — probably doable in a few minutes to a couple of hours.
- marked *Project* — probably a lot of work, but almost certainly doable.
- marked *Research Project* — almost certainly a lot of work, but hopefully doable.

We make no claims as to the relevance of any of these problems; we hope that some readers will find some of them enlightening, interesting, or perhaps even useful. Ideas, hints, and partial or complete solutions to a number of the problems can be found in Chapter A.

There are also a few questions on formal language that were not answered easily in the existing literature but have some importance to parsing. These have been marked accordingly in the problem sections.

Annotated Bibliography

For the first edition, we, the authors, read and summarized all papers on parsing that we could lay our hands on. Seventeen years later, with the increase in publications and easier access thanks to the Internet, that is no longer possible, much to our chagrin. In the first edition we included all relevant summaries. Again that is not possible now, since doing so would have greatly exceeded the number of pages allotted to this book. The printed version of this second edition includes only those references to the literature and their summaries that are actually referred to in this book. The complete bibliography with summaries as far as available can be found on the web site of this book; it includes its own authors index and subject index. This setup also allows us to list without hesitation technical reports and other material of possibly low accessibility. Often references to sections from Chapter 18 refer to the Web version of those sections; attention is drawn to this by calling them “(Web)Sections”.

We do not supply URLs in this book, for two reasons: they are ephemeral and may be incorrect next year, tomorrow, or even before the book is printed; and, especially for software, better URLs may be available by the time you read this book. The best URL is a few well-chosen search terms submitted to a good Web search engine.

Even in the last ten years we have seen a number of Ph.D theses written in languages other than English, specifically German, French, Spanish and Estonian. This choice of language has the regrettable but predictable consequence that their contents have been left out of the main stream of science. This is a loss, both to the authors and to the scientific community. Whether we like it or not, English is the de facto standard language of present-day science. The time that a scientifically in-

terested gentleman of leisure could be expected to read French, German, English, Greek, Latin and a tad of Sanskrit is 150 years in the past; today, students and scientists need the room in their heads and the time in their schedules for the vastly increased amount of knowledge. Although we, the authors, can still read most (but not all) of the above languages and have done our best to represent the contents of the non-English theses adequately, this will not suffice to give them the international attention they deserve.

The Future of Parsing, aka The Crystal Ball

If there will ever be a third edition of this book, we expect it to be substantially thinner (except for the bibliography section!). The reason is that the more parsing algorithms one studies the more they seem similar, and there seems to be great opportunity for unification. Basically almost all parsing is done by top-down search with left-recursion protection; this is true even for traditional bottom-up techniques like LR(1), where the top-down search is built into the LR(1) parse tables. In this respect it is significant that Earley's method is classified as top-down by some and as bottom-up by others. The general memoizing mechanism of tabular parsing takes the exponential sting out of the search. And it seems likely that transforming the usual depth-first search into breadth-first search will yield many of the generalized deterministic algorithms; in this respect we point to Sikkel's Ph.D thesis [158]. Together this seems to cover almost all algorithms in this book, including parsing by intersection. Pure bottom-up parsers without a top-down component are rare and not very powerful.

So in the theoretical future of parsing we see considerable simplification through unification of algorithms; the role that parsing by intersection can play in this is not clear. The simplification does not seem to extend to formal languages: it is still as difficult to prove the intuitively obvious fact that all LL(1) grammars are LR(1) as it was 35 years ago.

The practical future of parsing may lie in advanced pattern recognition, in addition to its traditional tasks; the practical contributions of parsing by intersection are again not clear.

Amsterdam, Amstelveen
June 2007

Dick Grune
Ceriel J.H. Jacobs

Parsing Techniques

A Practical Guide

Grune, D.; Jacobs, C.J.H.

2008, XXIV, 662 p., Hardcover

ISBN: 978-0-387-20248-8