B NOTES ON LANGUAGES AND ALGORITHMS

B.1 DEFINING LANGUAGES WITH BACKUS–NAUR FORM (BNF)

In this book, we define several languages, including the languages of propositional logic (page 166), first-order logic (page 187), and a subset of English (page 670). A formal language is defined as a set of strings where each string is a sequence of symbols. All the languages we are interested in consist of an infinite set of strings, so we need a concise way to characterize the set. We do that with a **grammar**. We have chosen to write our grammars in a formalism called **Backus–Naur form**, or **BNF**. There are four components to a BNF grammar:

BACKUS–NAUR FORM

BNF

TERMINAL SYMBOLS

- A set of **terminal symbols**. These are the symbols or words that make up the strings of the language. They could be letters (**A, B, C,** ...) or words (**a, aardvark, abacus,** ...) for English. For the language of arithmetic, the set of symbols is

  $$\{\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}, \mathbf{7}, \mathbf{8}, \mathbf{9}, \mathbf{+}, \mathbf{-}, \div, \times, \mathbf{(}, \mathbf{)}\}$$

NONTERMINAL SYMBOLS

- A set of **nonterminal symbols** that categorize subphrases of the language. For example, the nonterminal symbol *NounPhrase* in English denotes an infinite set of strings including "you" and "the big slobbery dog."

START SYMBOL

- A **start symbol**, which is the nonterminal symbol that denotes the complete strings of the language. In English, this is *Sentence*; for arithmetic, it might be *Exp*.
- A set of **rewrite rules** or **productions** of the form $LHS \rightarrow RHS$, where *LHS* is a nonterminal, and *RHS* is a sequence of zero or more symbols (either terminal or nonterminal).

A rewrite rule of the form

$$Digit \rightarrow \mathbf{7}$$

means that anytime we see the string consisting of the lone symbol 7, we can categorize it as a *Digit*. A rule of the form

$$Sentence \rightarrow NounPhrase\ VerbPhrase$$

means that whenever we have two strings categorized as a *NounPhrase* and a *VerbPhrase*, we can append them together and categorize the result as a *Sentence*. As an abbreviation, the symbol |