is a command, but in natural language the hearer has to determine if an utterance is a command, question, statement, promise, or whatever. Formal language theory provides a framework within which we can address these more difficult problems, but it does not answer them on its own. In the remainder of this section, we define a formal language for our English subset, $\mathcal{E}_0$.

### The Lexicon of $\mathcal{E}_0$

LEXICON

The first step in defining a grammar is to define a **lexicon**, or list of allowable vocabulary words. The words are grouped into the categories or parts of speech familiar to dictionary users: nouns, pronouns, and names to denote things, verbs to denote events, adjectives to modify nouns, and adverbs to modify verbs. Categories that may be less familiar to some readers are **articles** (such as *the*), **prepositions** (*in*), and **conjunctions** (*and*).[3] Figure 22.5 shows a small lexicon.

ARTICLES
PREPOSITIONS
CONJUNCTIONS

Each of the categories ends in ... to indicate that there are other words in the category. However, it should be noted that there are two distinct reasons for the missing words. For nouns, verbs, adjectives, and adverbs, it is in principle infeasible to list them all. Not only are there thousands or tens of thousands of members in each class, but new ones are constantly being added. For example, "fax" is now a very common noun and verb, but it was coined only a few years ago. These four categories are called **open classes**. The other categories (pronoun, article, preposition, and conjunction) are called **closed classes**. They have a small number of words (a few to a few dozen) that could in principle be enumerated. Closed classes change over the course of centuries, not months. For example, "thee" and "thou" were commonly used pronouns in the seventeenth century, were on the decline in the nineteenth, and are seen today only in poetry and regional dialects.

OPEN CLASSES
CLOSED CLASSES

### The Grammar of $\mathcal{E}_0$

The next step is to combine the words into phrases. We will use five nonterminal symbols to define the different kinds of phrases: sentence (*S*), noun phrase (*NP*), verb phrase (*VP*), prepositional phrase (*PP*), and relative clause (*RelClause*).[4] Figure 22.6 shows a grammar for $\mathcal{E}_0$ with an example for each rewrite rule.

## 22.4 SYNTACTIC ANALYSIS (PARSING)

There are many algorithms for **parsing**—recovering the phrase structure of an utterance, given a grammar. In Figure 22.7, we give a very simple algorithm that nondeterministically chooses one possible parse tree, if one exists. It treats the list of words as a **parse forest**: an ordered list of parse trees. On each step through the main loop, it finds some subsequence of elements in the

PARSE FOREST

[3] The term **conjunction** here means something that joins two phrases together. It addition to *and*, it includes *but, since, while, or*, and *because*. Do not be confused by the fact that *or* is logically a disjunction but syntactically a conjunction.

[4] A relative clause follows and modifies a noun phrase. It consists of a relative pronoun (such as "who" or "that") followed by a verb phrase (and sometimes a whole sentence). An example of a relative clause is *that gave me the gold* in "The agent *that gave me the gold* is in 2,2."