quency (from a sample of 1,000,000 words) in the given language is 150. Substituting these values into the preceding formula, we obtain the following:

$$w = \frac{120 \times 1000000}{50000 \times 150} = 16.$$

If we want our dictionary to contain the 1000 most important terms, we can easily arrange this after calculating all word weights of the collection of documents. Note that this method prevents us from selecting link words and commonly used words because these do not acquire high weights.

We have now considered the main approaches to the automatic construction of descriptor dictionaries. This problem seemed to be one of the most important problems in developing IRL. However, recent works in the area show the aspiration of developers to consider all words in the collection of documents as descriptors. This tendency becomes especially evident when dealing with new IR systems. In other words, in this case there is no need to create special descriptor dictionaries. This new angle to the approach taken by Taube has become practically feasible (and useful) due to the explosive development of computer technologies rather than to theoretical discoveries in the field of information science. The speed of modern computers and the size of internal memory have caused the revision of several theoretical principles that until recently had been regarded as unshakable. Also, the descriptor dictionary is not the only example of the influence of computer technology on IR system development. This fact will be quite apparent in the subsequent chapters.

Remember that the use of dictionaries similar to those applied to the natural language does not imply the use of the natural language itself, because the retrieval in the system does not utilize either grammatical or semantic components of the natural language. In addition, in some cases so called *nonimportant words* (link-verbs, commonly used words, etc.) are rejected from the collection of documents. Dictionaries of nonimportant words (and not the descriptor dictionaries) are compiled for this purpose. Obviously, to select nonimportant words, one can use the automatic methods for the selection of terms described earlier (with minor modifications). For instance, after weights of all terms in documents are calculated, one can select those with minimal weights, that is, those that are least representative for a given collection of documents.

Later in the book, when discussing the process of automatically indexing documents, we will consider the construction of document profiles both for systems using only descriptor dictionaries and for those utilizing dictionaries of nonimportant words (these dictionaries are also sometimes called dictionaries of stop-words).

Now, having considered the lexical component of IRL, we can proceed with the description of its grammatical and semantic components.

5.6

Semantic and Grammatical Components of IRL

It should be noted that, excluding cases when developers work with a free text search (see, for example, Lancaster, 1979), the compilation of descriptor dictionaries is virtually the only significant activity performed when creating IRL for a concrete system. Indeed, as a rule, system developers do not create semantical or grammatical components of IRL. The reasons are similar to those given by people who develop new models of cars without introducing principally new internal-combustion engines; in developing new models, such experts base their work on well-known principles accepted through many years of practice. In essence, the same process occurs during the creation of IR systems. The majority of investigators base their work on IRL principles formulated decades ago. This means that they consider semantic and grammatical components of IRL as something given. However, what is given are not universal rules but different alternatives for developers (although very few). Moreover, these alternatives are based on the ideas of Mooers and Taube. It seems that we can refer here to a family of descriptor languages because they have different grammatical or semantical components. Thus, as one can see from the analysis of existing systems, given a choice of a descriptor dictionary, developers choose rules for grammatical and semantical components of IRL that they think will provide for the highest quality information retrieval. Obviously, no universal opinion exists concerning the quality of existing rules. However, those rules chosen (and used) by the majority of developers might be considered as most appropriate. But what rules do developers actually use and what are the most popular among them? To answer this question, we will discuss the various existing rules; that is, we will describe the main existing rules for the IRL components under consideration.

The semantical and grammatical components of IRL are used to represent both the objects of retrieval and the requirements for them. In other words, grammar and semantics are necessary for writing (composing) phrases (texts) in IR language. During retrieval, these phrases should represent documents and search requests. Obviously, these phrases do not represent physical properties of documents and search requests, but instead these phrases represent their content and meaning. Another obvious fact is that phrases representing meanings in IRL are intended not for reading, but for finding information in IR systems. Therefore both grammatical and semantical rules are introduced in a form that allows them to be used directly during information retrieval. In essence, it is the selection criterion that determines the usage of specific grammatical rules in writing phrases in IRL. However, as was pointed out earlier, to introduce a selection criterion itself one needs the semantical component of IRL. Indeed, to find re-