Regarding the first question, some investigators hold that IRL should be created by linguists, whereas others suggest that the best quality will be provided by experts in the areas for which IR systems are created. For instance, a geological descriptor dictionary should be compiled by geologists, whereas a dictionary on computer science is the business of experts in this area. Some investigators reason that IRL should be created by the developers of IR systems, that is, by experts in the field of information science. One would admit that there are a lot of arguments in favor of all these points of view, which is probably why, in practice, IRLs are created by experts from all the professions mentioned, and in many cases they are created by groups that include experts from all three categories.

Somewhat more difficult questions arise when resolving the other problems mentioned. In essence, they all question *how* to create IRL? Traditionally, experts have created descriptor dictionaries "manually" (or, as one would say, "on an intellectual level") using their own vision of the problem, knowledge, and comprehension. This led (since the late 1950s) to the appearance of a considerable bulk of publications describing some well-intentioned advice and wishes, usually called methodological recommendations. They were based either on the personal experience of authors or on speculative concepts that seemed reasonable to them (see, for example Broadhurst, 1962; Francisco, 1956; and "The Uniterm System," 1955). Although some of these recommendations were rather useful, they failed to play even a third-rate role in the creation of IRL. This can be illustrated by the following example. Many recommendations have been offered on how to translate from one natural language to another. However, when one does not know a language, no recommendations will help. On the other hand, brilliant knowledge of two languages makes translation possible without any recommendations.

The situation with descriptor dictionaries is quite similar. Developers need to know very well what a descriptor language is. But that is not enough. "Manual" creation of IRL is based on the developer's personal decision to form each descriptor. Therefore, a developer's success essentially depends on his or her knowledge of the subject matter of the document collection and on the knowledge of all features and fine points concerning the further use of the dictionary in a concrete IR system. This by no means implies that no methodologies are needed. They are needed. For instance, most of the methodologies for creating descriptor dictionaries recommend the use of concise dictionaries, encyclopedias, terminological manuals, thesauruses of natural language, and so on. Although this advice is obviously useful, the success of the personal creation of the dictionary nevertheless depends on the skills of its developer. That is why we will not concentrate on either manual methods or manual methodologies. Because manual methods were labor intensive, qualitatively unstable, and personality dependent, developers searched for ways to compile dictionaries automatically. Therefore, in the following section we shall confine ourselves to

the consideration of methods used for the automatic selection of terms for descriptor IRLs.

5.5 Automatic Methods of Descriptor Dictionary Compilation

Actually, descriptor dictionaries were compiled manually not because the developers considered manual methods superior to automatic methods. There were just no ideas at the time to explain how to automate the process. Nevertheless, the majority of investigators did understand the advantages of automation. G. Salton, one of the most distinguished creators of IR systems, indicated that

> in normal conditions the creation of a dictionary for a given subject area requires advanced skills, persistence and intensive will. . . . Since the volume of the problem is large, frequently the whole committee is organized to resolve arguable points and, eventually, to create a dictionary. Such a dictionary created by a committee may not satisfy anyone despite great efforts to create it.
>
> Obviously, if one follows this scheme of compiling dictionaries, any profit resulting from the automated retrieval will be immediately lost due to the complexity of the dictionary creation. That is why this situation has encouraged many attempts either to create dictionaries completely automatically or, at least, to use methods more effective than the work of a committee. Any sufficiently universal method of dictionary creation not only saves time and reduces costs, but also provides essentially more freedom in choosing types of retrieval procedures to implement. (Salton, 1968)

Virtually all ideas of the automatic choice of words for descriptor dictionaries follow from certain statistical characteristics of texts composed in various natural languages. An intensive study of the statistical laws of languages started almost simultaneously with the creation of computers. Even in the early 1950s, linguists paid great attention to the possibilities provided by computers. For instance, for various characters in an alphabet, frequencies of use were calculated using large text collections. These frequencies were calculated for different alphabets and languages. Frequencies of the joint appearance of characters in words (for two characters, three characters, and so forth) were also calculated. As an example, in Figure 5.1 we present part of the character frequency table for texts in Russian.

Frequency characteristics of characters have found numerous applications in information encoding and transfer. However, for problems of information retrieval, the calculation of word (rather than character) occurrence frequencies turned out to be even more important. A fragment of the word list with the relative word frequencies is given in Figure 5.2 as an example. Such lists were