creating abstracts are frequently considered similar tasks. However, there is a principal difference between indexing and abstracts: indexing is the process of translation from one language to another, whereas the creation of abstracts does not require translation. In addition, the result of indexing is not intended to be read, whereas that is the only objective of an abstract. The result of indexing is written in an artificial language, and it does not need the semantic components of natural language, whereas the abstract is always represented in natural language and, consequently, must operate with the semantic components of natural language. Thus, we would say that the specificity of indexing is closer to object classification than to creating abstracts.

The second direction proposes a completely different approach to determining the importance of terms contained in a document being indexed. In this case, when the document is analyzed by computer, the system determines how well a certain term represents the thematic area of the IR system is intended for, rather than the meaning of the document (its main meaning content). The term may be very important for the meaning of the document being indexed, but it will not be used as a descriptor (not included in the document profile). On the other hand, a certain term may concern second-priority issues that are not covered in detail in the document, but may be important for the thematic area covered by the system. Then the term will be regarded as a descriptor and will be included into the document profile.

In essence, the automatic indexing aimed at determining the importance of a term with respect to the thematic area of the system is significantly simplified. The point is that the selection of important terms (descriptors) is performed before indexing, during the selection of IRL lexical units (that is, during the creation of the system vocabulary). In other words, by definition (see Chapter 5) only the terms that most completely reflect the terminology used in the framework of the thematic area of the system are included in the descriptor dictionary. That is why the presence of fixed descriptor vocabularies is necessary for systems where indexing is performed according to the second approach. In essence, the document indexing procedure in these systems is confined to finding terms included in conditional equivalence classes of descriptors. These descriptors make up the document profile. This approach to the compilation of the document profile clearly provides "equality" (equal importance) to all descriptors with respect to the representation of the document meaning. This means that this approach is mainly intended for the Boolean search, because one can regard all descriptors as having no weights (or they all have equal weights). Indeed, this approach has practical implementations. It is worth noting that the developers of automatic indexing subsystems have found it attractive not only because the idea of its practical implementation is simple but because of other factors as well. The point is that methods based on this approach have demonstrated higher retrieval quality than those using the first approach. Although it is quite evident that the data on the comparative efficiency of indexing methods given in current

publications are valid only for specific constructions of specific systems, these data have some general validity, at least for systems with similar construction.

The first realistic approaches to automatic indexing can be found in the late 1950s (Luhn, 1957; Luhn, 1958). H. P. Luhn, a well-known researcher at IBM, is probably the father of automatic indexing. It is noteworthy that he started both of the approaches mentioned earlier. For the first approach, he suggested that statistical characteristics of text should be used for indexing, that is, for determining descriptors that are most significant for representing meaning of the document being indexed. Additionally, he suggested a simplified but rather concrete method for this kind of indexing. With regard to the second approach, it was Luhn who adapted the idea of word-to-word translation with the aid of a dictionary for the purposes of indexing. Previously this idea was used in automatic translation of texts from one natural language to another. (In the late 1930s, the translation of texts from, say, English to French with the aid of an English-French dictionary was suggested.) Within the framework of the second approach, Luhn suggested the use of a precompiled descriptor dictionary for the automatic indexing of documents based on a vocabulary created for the IR system. The indexing procedure itself was merely a comparison of each word of the document with the words (descriptors) contained in the precompiled descriptor dictionary. Luhn reasoned that if the word from the text coincides with a word (descriptor) from the vocabulary, then this descriptor should be introduced into the document profile.

As indicated previously, this particular idea has, in essence, served as a base for the majority of contemporary automatic indexing subsystems that are practically implemented and successfully operated. Because functioning IR systems are, as a rule, represented by the systems utilizing the Boolean search, such methods of automatic indexing are of particular interest to us here. Therefore, in the following discussion, we pay special attention to the indexing process using a dictionary and its practical implementation.

6.4 Some Methods of Text Analysis for Automatic Indexing Using a Dictionary

It seems that the idea of automatic indexing using a dictionary could immediately allow us to describe some type of functioning algorithm. However, it will be useful to first describe the technical difficulties encountered in designing such an algorithm.

To begin with, note that people normally (for instance, when reading) comprehend the meaning of a text without analyzing the meaning of each word. In the majority of cases, we know all the words of the text very well and we