As the preceding discussion shows, using the algorithm to help the IR system recognize word combinations is not too difficult to implement. In our opinion, such algorithms can be useful, particularly when the system dictionary contains a considerable number of word combinations. Although, as we have mentioned, several well-known experiments indicate the absence of significant improvements resulting from recognizing word combinations, it seems that this can be explained by the rather coarse methods of both implementation and evaluation of retrieval.

Another measure, probably the most frequently undertaken in automatic indexing, is to use lists of nonimportant terms. They are frequently called the lists of stop words (see Chapter 5). A portion of such a list is presented in Figure 6.3. This list was compiled at Fordham University by students in the information retrieval systems course.

Although the use of such lists in indexing with the aid of the descriptor dictionary is not necessary (their use does not influence the indexing result), it is very desirable, mainly for two reasons. First, the removal of stop words from the document being indexed (in the framework of the statistics of the given system) can enhance the performance of the entire automatic indexing process. Second, the use of stop word lists is important in a number of methods used for the automatic monitoring of the lexical vocabulary of the thematic area that the indexed document represents. Such monitoring is required to make a direct change (correction) to the descriptor dictionary used for indexing. In other words, this makes the system speak the same language that is used by the authors of the documents being indexed. We next consider in detail each of the two reasons for using stop word lists.

a	ago	an	at
abound	all	and	be
about	allow	another	began
above	almost	any	became
across	alone	anyhow	because
admit	along	anyone	become
afere	already	anything	becomes
aforetime	also	anyway	becoming
after	although	anywhere	been
afterwards	always	are	before
again	among	around	beforehand
against	amongst	as	begin

Figure 6.3
Excerpt from a typical stop list.

We mentioned that the use of stop word lists can enhance indexing performance. To be more accurate, stop lists can help to improve (reduce) indexing time. Let us consider how they can help to achieve this. In essence, the list of stop words not only consists of common words (their number in the English language is estimated at about 250) but also often contains words that are considered unimportant in the framework of the given IR system (at least by the creators of the descriptor dictionary) but that are frequently encountered in the documents being indexed. Common words make up about 50% of all words in document texts. Therefore, if all stop words that have been included (fixed) in the stop list are removed from the document, the indexed text will be substantially reduced. Thus, the time required to create the document profile is also reduced. This is particularly important when the automatic indexing algorithm involves morphological analysis (for example, stem separation), semantic analysis (determining meanings of homonyms), or syntactical analysis (recognizing word combinations), and especially their combination. In other words, the more procedures of text analysis involved in automatic text indexing, the more essential the positive effect of the stop list becomes. Theoretically, no positive effect may occur if stop lists are too large (due to the vast number of comparisons for each word in the text). However, we do not know practical examples of this kind.

Another reason for the use of stop lists is the correction of the descriptor dictionary itself. In essence, the necessity of such a correction is dictated by the dynamic nature of natural language, by its constant change. Changes are most noticeable in scientific literature, because in the course of scientific activities new effects and laws, as well as new objects and phenomena, are discovered. Naturally, discoveries like these provide new terminology, often altering the former terminology by attributing new meanings to old terms and sometimes just rejecting terms that were previously used. It is evident that IR system descriptor vocabularies should reflect the current terminology in the area it is created to assist. This, in turn, should improve the quality both of indexing and of retrieval in general. Although the automatic correction of descriptor vocabularies has only an indirect relation to automatic indexing, it should nevertheless be performed in the course of indexing, at least partially. Mainly this is determined by the fact that automatic indexing includes the total examination of all words contained in each document text being indexed. During indexing assisted by a stop list, the appearance of a new, somewhat accepted term (i.e., a term that is encountered in documents of the collection but is not included either on the stop list or in the descriptor dictionary) cannot be missed. Such a control allows the IR system to correct not only the descriptor dictionary but the stop list as