NSDL Annual Report 2005: Grade Level Analysis of Eisenhower National Clearinghouse documents

Peter Shin, Tony Fountain, Reagan Moore
San Diego Supercomputer Center, UCSD

Data provided by Judy Ridgway
The Eisenhower National Clearinghouse

As students advance to higher grade levels, they learn new words. The documents intended for upper grade levels will contain more advanced vocabularies, reflecting the assumed aptitude level of the intended audience. In this study, we first classified all the words in a pre-labeled document collection into various grade level categories. We then calculated the distribution of words from each grade level for all the documents. The eventual goal of our study is to build a system that automatically assigns the appropriate grade level label to each document in the NSDL repository. This will allow the educators to search more easily for material appropriate to specific audiences.

The available dataset for this study comes from the Eisenhower National Clearinghouse. This dataset contains a total of 8,417 documents with labels specifying the intended grade levels. Figure 1 displays the number of documents that are appropriate for the various grade levels. Since many of the documents cover more than one grade level, summing up such documents correspond to a total larger than the number of unique documents (over 50,000 vs. 8,417).

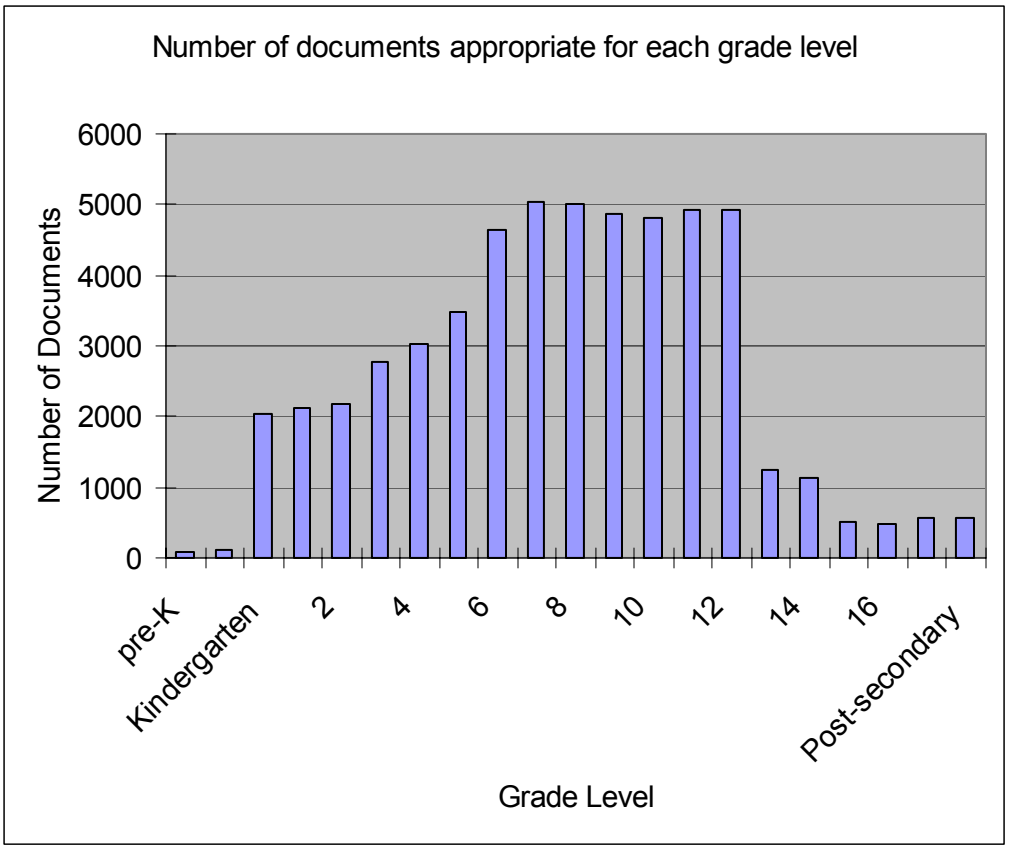

Figure 1

These "multiple count" documents are further explored in Figure 2, which shows the distribution of the number of grade levels covered by each document. Documents that span a large number of grade levels are less useful in identifying grade level-specific words. Thus, by considering the documents with narrow span of grade levels, one can more accurately build a catalog of words that are specific to different grade levels.

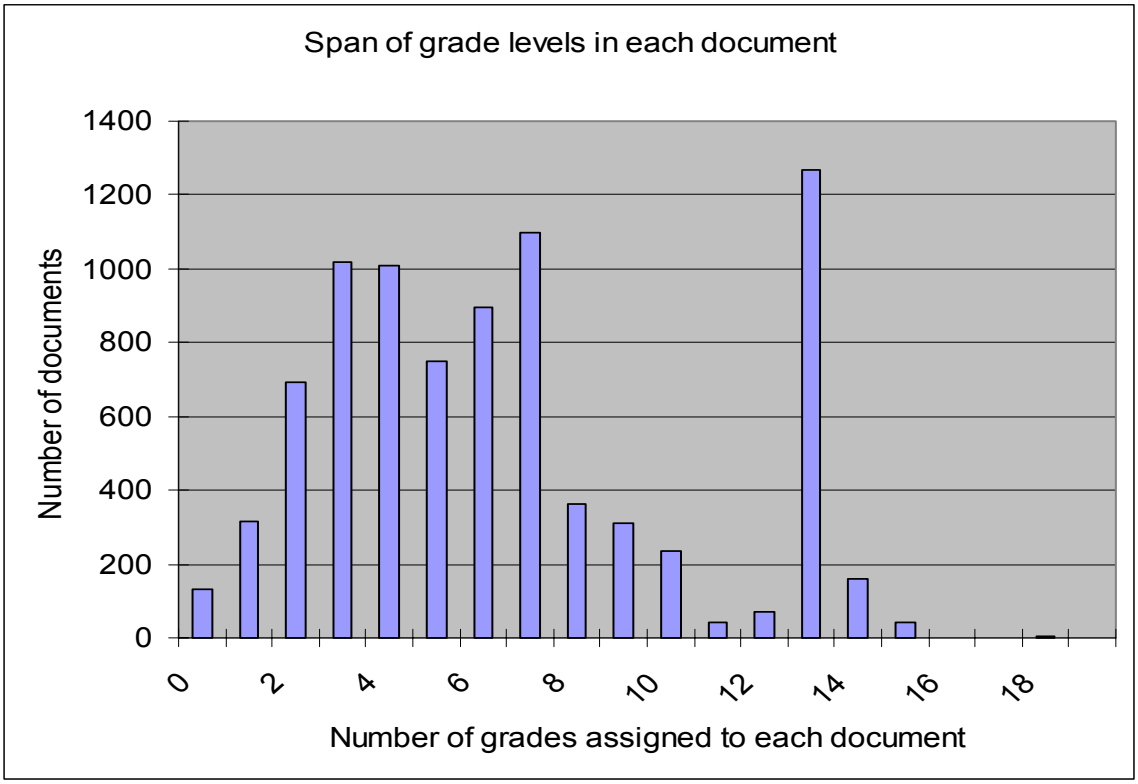

Figure 2

As shown in figure 2, there are 315 documents with only one grade level span. Figure 3 shows the distribution of documents that are labeled as appropriate for a single grade level. More than two thirds of these documents are in the highest grade level region (higher than grade 12). The small number of document prohibits building an extensive word list.

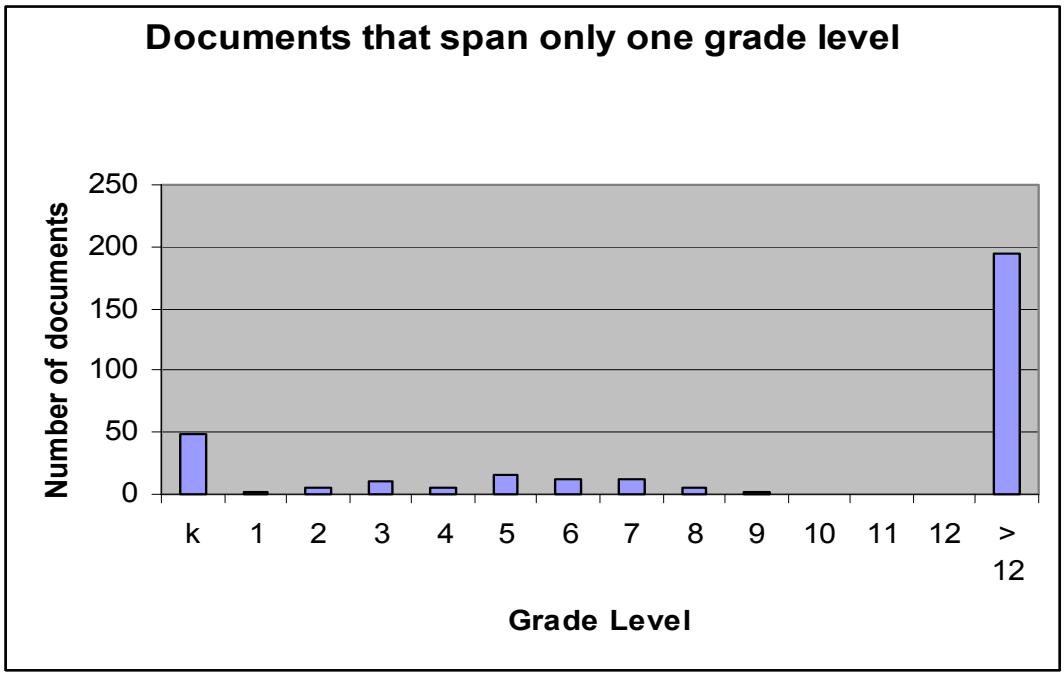

Figure 3.

One of the NSDL text mining goal is to map the NSDL material to the American Association for the Advancement of Sciences (AAAS) strand maps. The AAAS breaks down the grade levels in the following range: k-2, 3-5, 6-8, and 9-12. Using this convention, 1353 documents matched one of the AAAS grade ranges. Figure 4 shows the distribution of documents that fall into each AAAS category, and figure 5 shows the number of words found in each grade category.

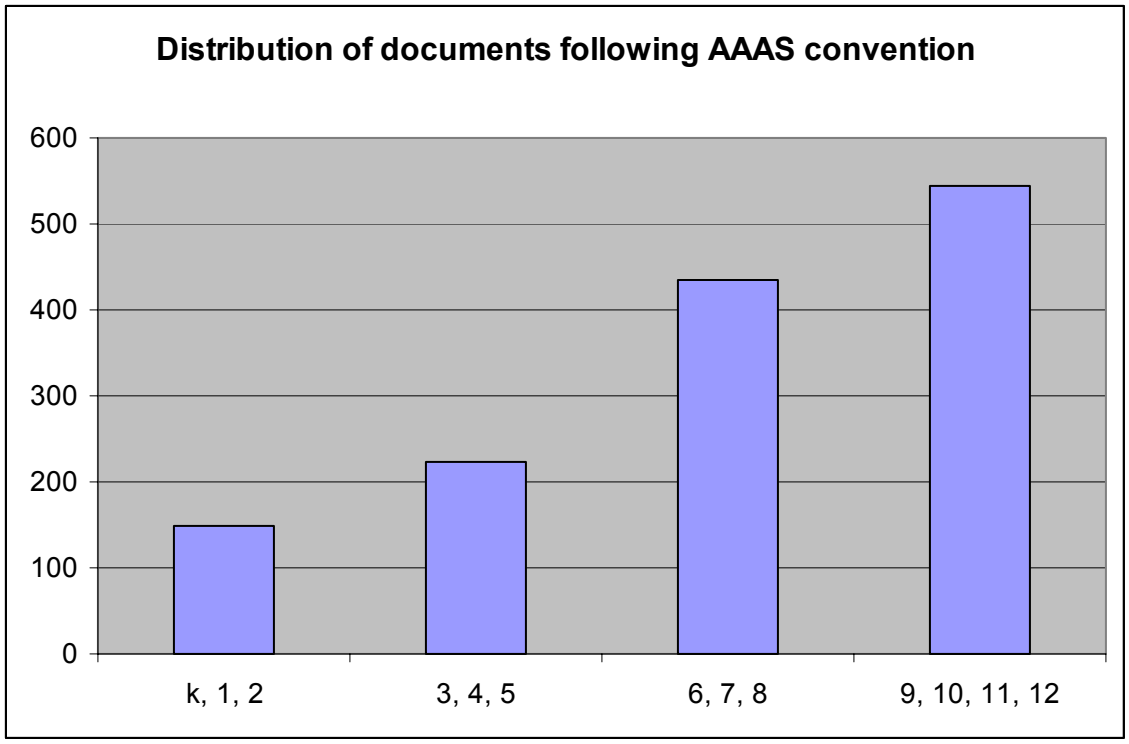

Figure 4.

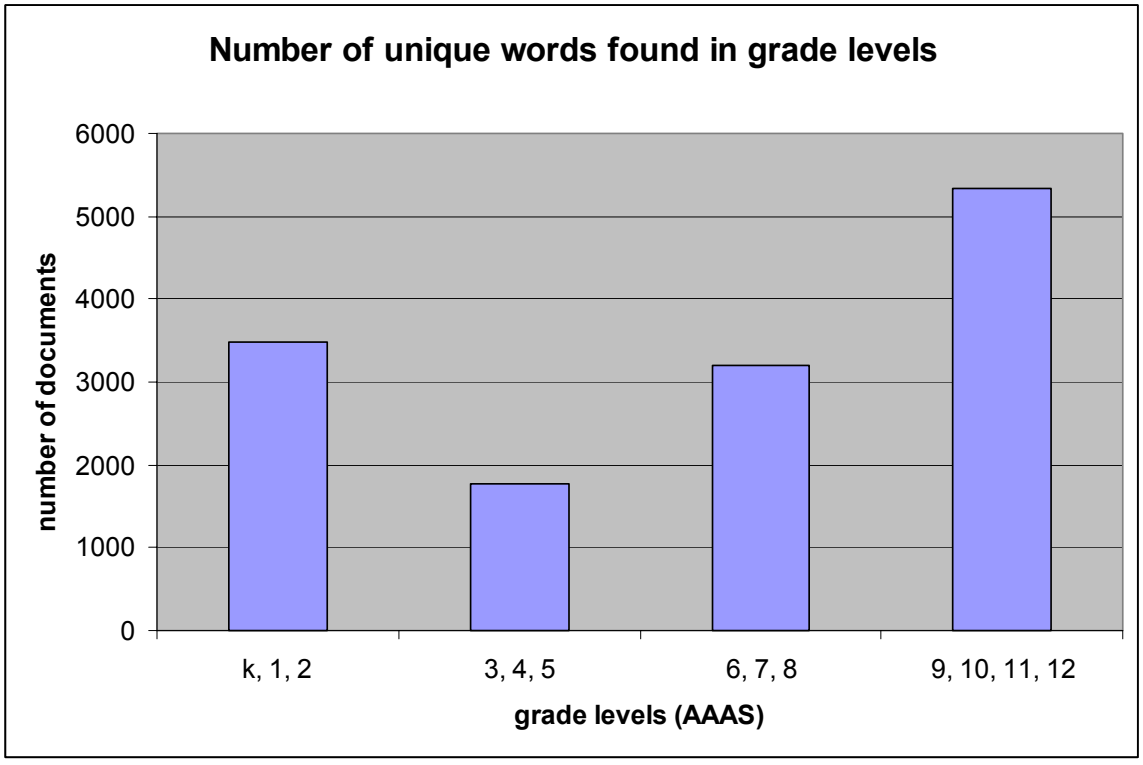

Figure 5.

In looking at these figures, it becomes apparent that a large number of documents in each category (Figure 4.) does not translate to more new words appearing in Figure 5. This is

mainly caused by the overlapping of many words within that category. Also, each document may contain a large portion of words from lower grade categories. This is further illustrated in the following pie graphs (Figure 6 through 9), which show the average distribution of various grade level words for a typical document in each document category.

In calculating these various distributions below, an interesting trend became apparent. Although the number of unique words in our constructed vocabulary increased with the number of documents in each category (Figure 4 and 5), on average, the percentage of the words in each document that were actually new decreased (Figure 6 through 9). For example, in the average k-2 document (Figure 6), 100 % of the words were considered new. In later grade levels, however, the percentage of new words gradually decreased. This can be seen most noticeably in the grade 9 through 12 documents, where on average, the actual percentage of new words on average was 10 %.

By combining the corresponding categories in Figure 4 and Figure 6 through 9, the result of Figure 5 can be explained. Specifically, although the least percentage of the new unique terms introduced in each document occurs in the highest grade level (Figure 9), combining that percentage with the large number of documents in Figure 4, results in that category having the most number of new words in Figure 5.

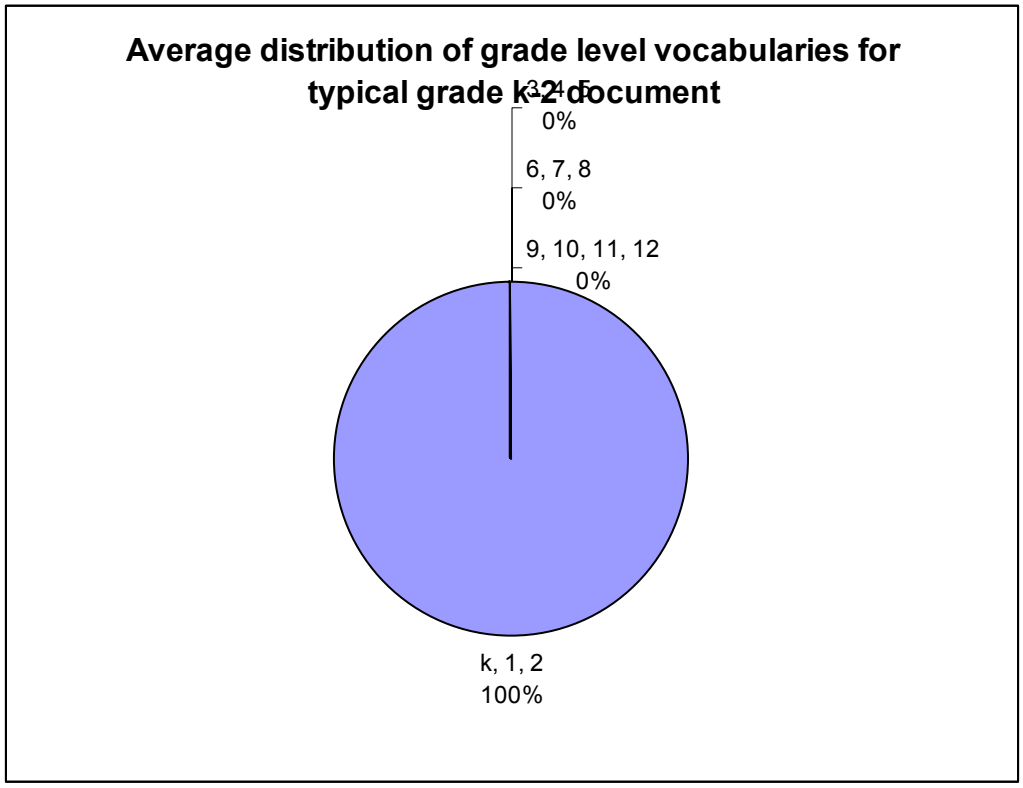

Figure 6.

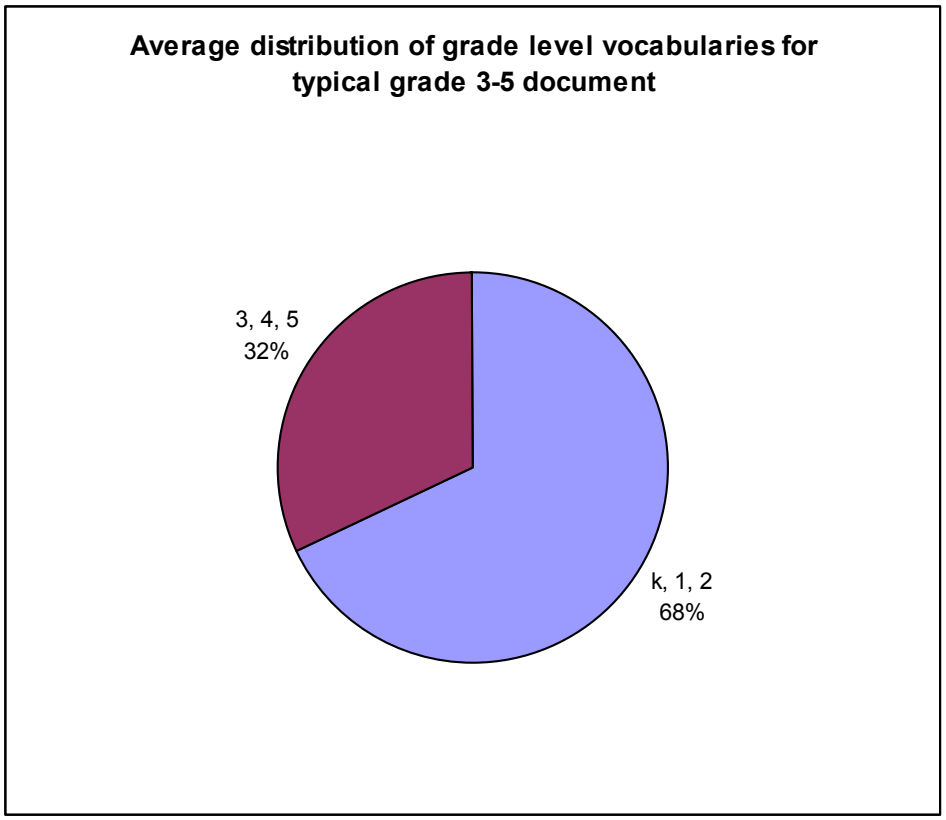

Figure 7.

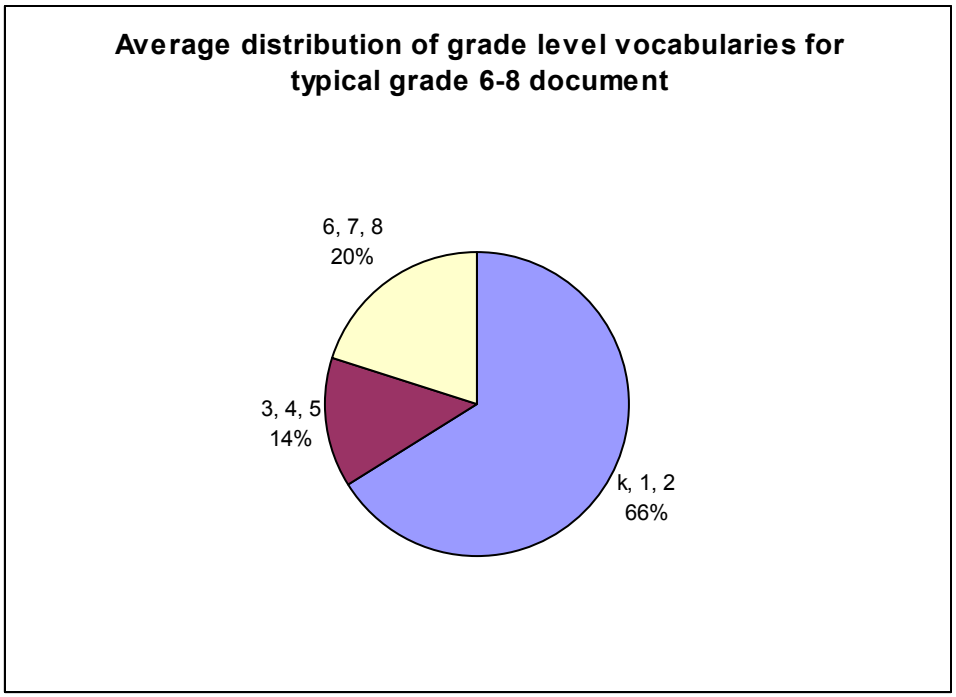

Figure 8.

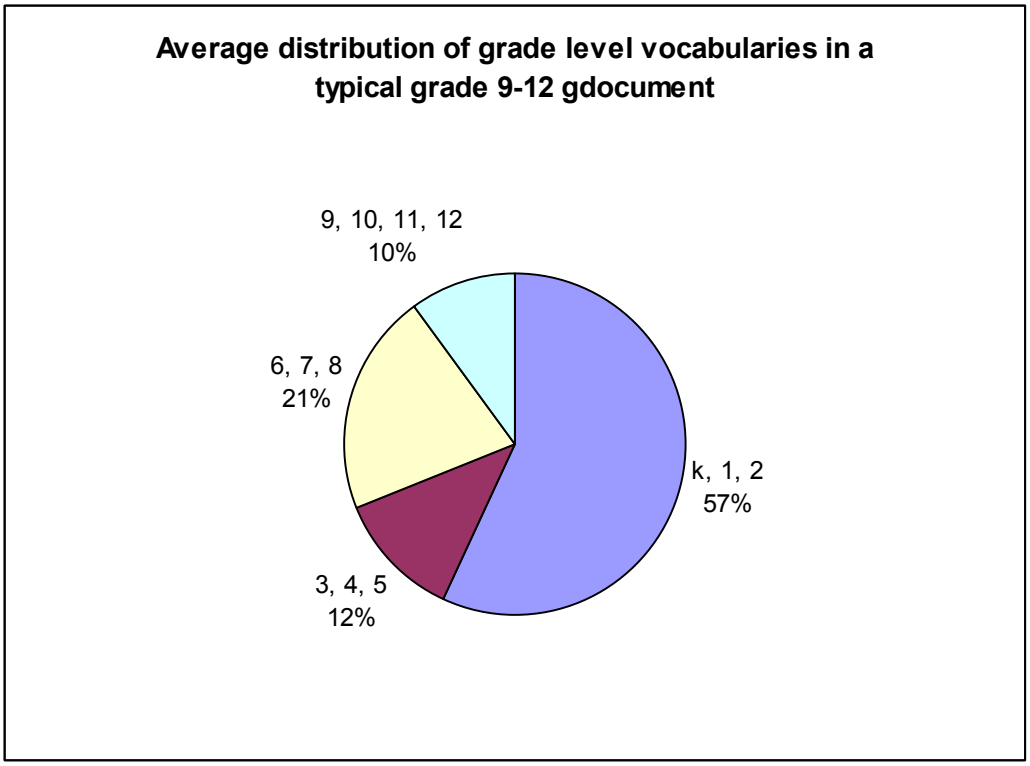

Figure 9.

These observations can be useful in constructing the desired grade level classifier for the NSDL collection. The percentage of highest grade level words within a document can be a good indicator for the appropriate grade level of that document. Studying this parameter would be critical in building such an automatic classification system. If the document only contained one word that is in a higher grade level than the rest of the words, how should that document be classified? As a next step, this idea can be further explored using machine learning techniques that assign weights to all the terms in a document as a group.