The statistics view allows you to review word usage in the entire corpus (Full corpus statistics), in individual authors or works (Author statistics) or view the usage of a particular lemma (Lemma statistics).
This page includes a summary of the data included in the TLG corpus. Information is presented in the forms of diagrams and expandable menus.

Information included under the expandable menus is linked to Text search or the TLG Canon.
This page provides statistical information about a specific author or work.

The information is given through diagrams or expandable menus (left column).

Each word in each list is linked to Text Search.

Two icons at the bottom of each diagram allow you to enlarge the diagram or obtain a list of the most frequent lemmata in the selected author or work.
Important notes

- There are certain lemmata which we ignore when looking at the most frequent lemmata in a subcorpus. Stop-words are ignored in N-grams and Statistics pages. Milesian numerals are also ignored in lists of statistics.

- The total number of words listed in the Summary section is the raw count which includes non-Greek words. Lemmata, wordforms and total word counts--broken down by century--do not include non-Greek words.

- Unique occurrences vs. hapaxes
  
  It is easy for the TLG search engine to identify lemmata that are unique to a specific author in the corpus. The statistics display lists of those lemmata as unique occurrences.

  Unique occurrences of lemmata are known as “hapaxes”. We do not use the word hapax, because our corpus is not the same as the classical corpus against which hapax was defined. A word is a hapax in Homer if it was not used elsewhere in the classical literary canon. But the TLG is not limited to the classical literary canon: it also includes Homeric dictionaries, Homeric scholia, grammarians, and literary theorists, all of whom discussed the words of Homer at length. So while there are numerous Homeric hapaxes by the classical definition, only 14 lemmata in Homer are restricted to Homer in the TLG corpus.
Relative frequencies of lemma use, rather than raw counts, are particularly important when we are comparing usage across different centuries. Through accidents of scribal preference and cultural preservation, the quantity of preserved Greek literature varies widely by century. The corpus for A.D. 4 is around 18 million words; the corpus for 8 B.C. is 0.2 million words, which is 90 times smaller. So it is meaningless to say that there are more instances of καλός in 4 A.D. than B.C. 8 (17,547 vs. 402); with 90 times more words, we would expect any word to occur more often in iv A.D. than viii B.C.

One way of looking at this is by working out what the expected count of the word would be per century, if the word were distributed evenly across the TLG. καλός occurs 111,604 times in the 98 million words of the TLG. Extrapolating from that, it should occur 20,230 times in iv A.D., and 274 times in viii B.C. There are over 2,300 instances of καλός less than we would expect in iv A.D.; so we can say that καλός is underrepresented by 2300 instances for that century. There are 127 instances of καλός more than we would expect in viii B.C.; so we can say that καλός is overrepresented by 127 instances for that century.
This page provides statistical information about a particular word. The information is given through diagrams or expandable menus (left column).

Click on Geographic Distribution to view this information on the map.
Click on Geographic Distribution to view information about a specific lemma on the map.