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.

- 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.
Over-represented and Under-represented Lemmata

The word λέγω occurs more frequently than any other content word in Plutarch. This does not tell us much since λέγω is an extremely frequent word in any author. To go on to say that λέγω occurs 100 times in Plutarch, and just 7 times in Pindar, is not a meaningful comparison either, given that we have over a million words of Plutarch in the TLG, and just 28 thousand words of Pindar.

The more interesting question to ask about frequent lemmata in an author is, which are the lemmata that the author uses much more than expected. That helps us identify which lemmata are characteristic of the author.

To work out which lemmata an author uses more than expected, we look at the frequency of lemmata in the corpus overall. We extrapolate how frequent we expect the lemma to be in our given author, if he used it as frequently as the average across the corpus. If the author uses the lemma significantly more than the average, we can tell that it is characteristic of the author. For example: the lemma πολύς occurs 374,000 times in the TLG corpus, which corresponds to 3.82 times per 1000 words. There are just over a million words in Plutarch; so if Plutarch used the word with the same frequency as the TLG average, we would expect 3916 instances of the word. Instead, Plutarch uses the word 6578 times; so the word is overrepresented in Plutarch. By going through all the lemmata used by Plutarch, we can identify the lemmata most overrepresented in Plutarch — in a sense, the lemmata that are most distinctive to Plutarch.

The TLG is quite uneven as a corpus, with a wide variety of texts, and usage of words changed significantly across the centuries. So rather than ask whether a lemma is overrepresented in Plutarch, compared to the average across the entire TLG, a more sensible question to ask is, whether it is overrepresented compared to Plutarch’s contemporaries. To do this we look at how frequent the lemma is across all 1 AD authors, instead of how frequent it is across all of the TLG. For our example of πολύς, we find that it occurs 17,570 times among 1st c. AD, who account for 3.4 million words — in other words, it occurs 5.26 times per 1000 words. If Plutarch used the word with the same frequency as his average contemporary, we would expect 5374 instances of the word, instead of the 6578 times he actually uses. So the word is still overrepresented in Plutarch.
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.
Over-represented and under-represented lemmata per century

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.