N-grams are overlapping sequences of content words in text. They provide an efficient mechanism for identifying common passages between texts: by identifying sequences of two or three content words shared between two texts, we can quickly identify text passages in common.
N-gram comparison uses trigrams common between two selected works or authors, to identify common passages. The words in common between the two texts are highlighted. The comparison allows for small interpolations in the common passages; the words actually in common are highlighted, so un-highlighted words between two highlighted sections indicate interpolations (or deletions, if the apparent interpolation is in the original text).

The comparison is based on lemmata rather than wordforms, so it allows for differences of inflection between the two texts, as well as slight differences in word order (through unordered n-grams). There is a limit to the size of common passages identified between words, so that an extended passage common between two texts may be shown chopped up into several consecutive matches. If more lines of context are selected for the match, the maximum size of the match is increased, as is the maximum size of allowed interpolations.
Use of N-grams

The TLG website uses n-grams in three places/sections:

1) **N-Grams:**
   Compare two texts or two authors side by side

2) **Browse (Browse one text):**
   Enable N-grams and click on its phrase to see its attestation in the corpus

3) **Browse (Parallel Browing):**
   Select two passages and view their similarities
Browsing a text in n-gram mode (Browse, Links: N-Grams, Select: lemma or wordforms) highlights all the n-grams in a passage that also occur elsewhere in the TLG corpus. The n-grams can be based on wordforms, or on lemmata: if you select wordforms, only exact repetitions of the wordforms are reported, whereas lemmata allow for changes in inflection.

The n-grams highlighted include both trigrams and bigrams; the trigrams can span across sentences, but the bigrams cannot. (This helps contain the number of matches reported.)
The more frequently an n-gram occurs, the deeper its shade.

As you hover your mouse over the text, you will notice that the n-gram starting at the word you are pointing to is highlighted. If you click on the word, you get in the sidebar a list of all the instances in the TLG where that bigram or trigram occurs.
General rules

- Stop Words: N-grams are restricted to content words. We ignore stop-words that do not contribute much meaning, and which can distract from the underlying similarity of two texts.

- N-Gram order: We ignore the order of words within n-grams. This allows us to detect common passages between works, even if one of them swaps two content words.
  
  e.g.  
  πλείστον ἡμέρας τούτῳ μέρος  (Pl. Grg. 484e)  
  ἡμέρας πλείστον μέρος  (Arist. Rh. 1371b)

- Comparisons between authors in the Inter-textual Phrase comparison section are based on trigrams, and report matches containing a minimum of 2 trigrams and a maximum of 4 trigrams. That is, any matches in comparisons between authors report matches between 5 and 7 content words long.

- The shorter the match requested, the more irrelevant search results are displayed, and the longer the comparison takes to generate. We have found that the minimum match of 5-to-7 words is workable for inter phrasal search. On the other hand, if matches involve very short texts, critical similarities may be missed; for example, fragments consisting of just a title may not be matched against texts citing that title.

- Accordingly, if the Inter-textual Phrase comparison involves individual works rather than authors, and one of the two works is very short, matches use bigrams instead of trigrams, and require only one bigram for a match. This means that for very short works, a match need only contain two content words. (The criterion for switching to bigrams is that the work contains less than 10 trigrams occurring elsewhere in the work; this translates to 12-15 content words.)

- For Parallel Browsing, similarities are normally detected using a minimum of two trigrams. Again, if one of the two texts is very short, the comparison is made using a minimum of one bigram instead.

- Comparing Editions uses differences between individual word forms, beta escapes, and punctuation, rather than n-grams; so it captures finer distinctions between texts than n-grams do. Comparing editions still uses n-grams (with a minimum match of 2 trigrams) to align the two editions. The text in the old edition may need to be rearranged, to better match the new edition.