What Makes a Word Important?

Identifying Important Words in the Bible

Peter Venable, Ph.D. and Rick Brannan

Faithlife Corporation

April 11, 2019

Abstract

Logos 8 introduces a new feature that isolates important words in a Bible passage. But how is word importance determined? This talk will tell the story behind the development of the Important Words feature as well as discuss some of the technical details.

The Problem

In a word (ok, two words): Word Studies.

Word studies are very popular (well, in the New Testament, anyway). Classic volumes still in print (Robertson's Word Pictures, Vincent's Word Studies, Wuest's Word Studies) stop to examine words deemed worthy to comment upon. People do them as part of their exegetical method.

But how do people know which words are worth researching further for an in-depth study?

What Logos Bible Software Used to Do

Since circa 2006, way back to the Libronix Digital Library System, there was a feature (a "guide section") called "Interesting Words."

This was simply TF-IDF (**T**erm **F**requency / **I**nverse **D**ocument **F**requency) comparing the passage ("document") against the rest of the passages ("documents") in the Bible. It was useful because it was purely statistical which meant it could be run on any language Bible without any further annotation.

We really wanted to call it "Important Words" but all the feature really noted were words that were statistically interesting, not important. So we called it "Interesting Words." We made it back in the day when "word clouds" were the thing, as you can see:

And we always wanted to try and find something to replace it, because the real goal was to highlight words important to focus on while researching a passage.

And that was, for us anyway, state of the art for 12 years. Pretty pictures, useful for pasting into a bulletin or study guide, but not too useful for actually discerning the words in a passage that are key to properly understanding the passage.

The Interim

For most of the interim, life simply went on. We didn't have a good solution to determine "Important Words" so we didn't pursue anything.

Lemma in Passage

And then, in an at-the-time unrelated effort,[1] we decided that we should try and lemmatize all the Greek, Hebrew, and Aramaic strings we could find in commentaries because it would make the commentary material more accessible. After all, who doesn't want to be pointed to relevant discussions about particular instances of words in the passage being studied? This happened sometime in the Logos 6.x product cycle.

After we did that, we decided to try and analyze transliterated forms of Greek, Hebrew, and Aramaic as well.

This was (and is) a massive project. Consider the data points:[2]

- Total Commentaries: 7,055
- Commentaries with Original Language strings: 5,914
- Total Lemmas Analyzed: 8,716,453

I say it still is a massive project because we publish new commentaries all the time (Around 700 since we first published the Lemma in Passage data). We also update commentaries all the time. And when either of those two events happen, the underlying data needs to be updated as well. We release these updates every two weeks.

What the Lemma in Passage data does is locate the Hebrew, Aramaic, or Greek text in a resource. Then it splits the string into words and records the resource location of each word and processes it[3] to associate lemmas with the original languages. It means you can locate commentaries that have discussions about words in your passage. It's like having a lexicon, only the discussion about meaning, instead of being generalized across the lexicon's corpus, is applied to your current passage.

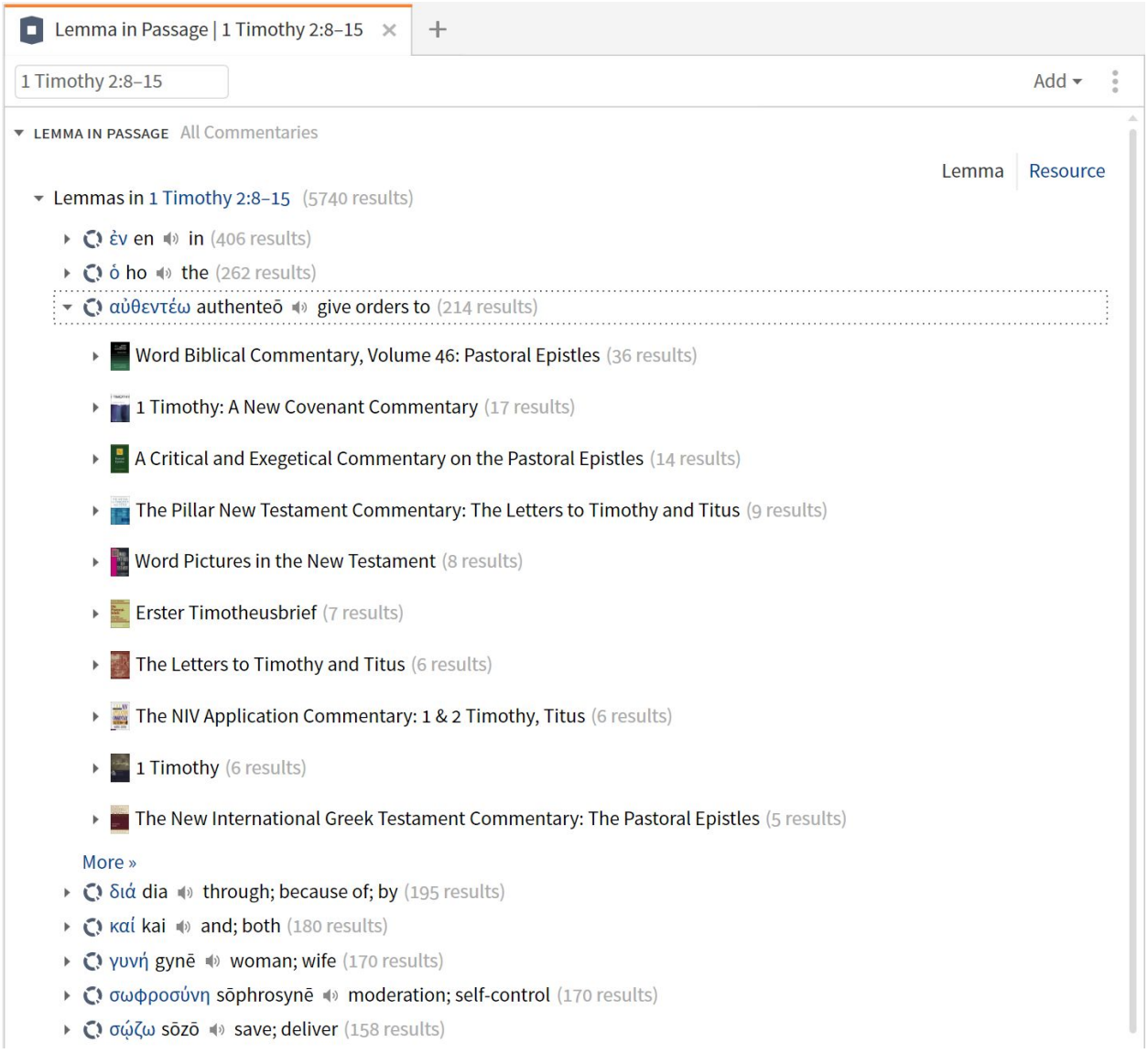
---

[1] It was in 2015, I believe. The git repo for the code that isolates Greek, Hebrew, and Aramaic in commentaries and lemmatizes it has an initial date of May 29, 2015 — about 2 weeks after the last BibleTech conference in May 2015.

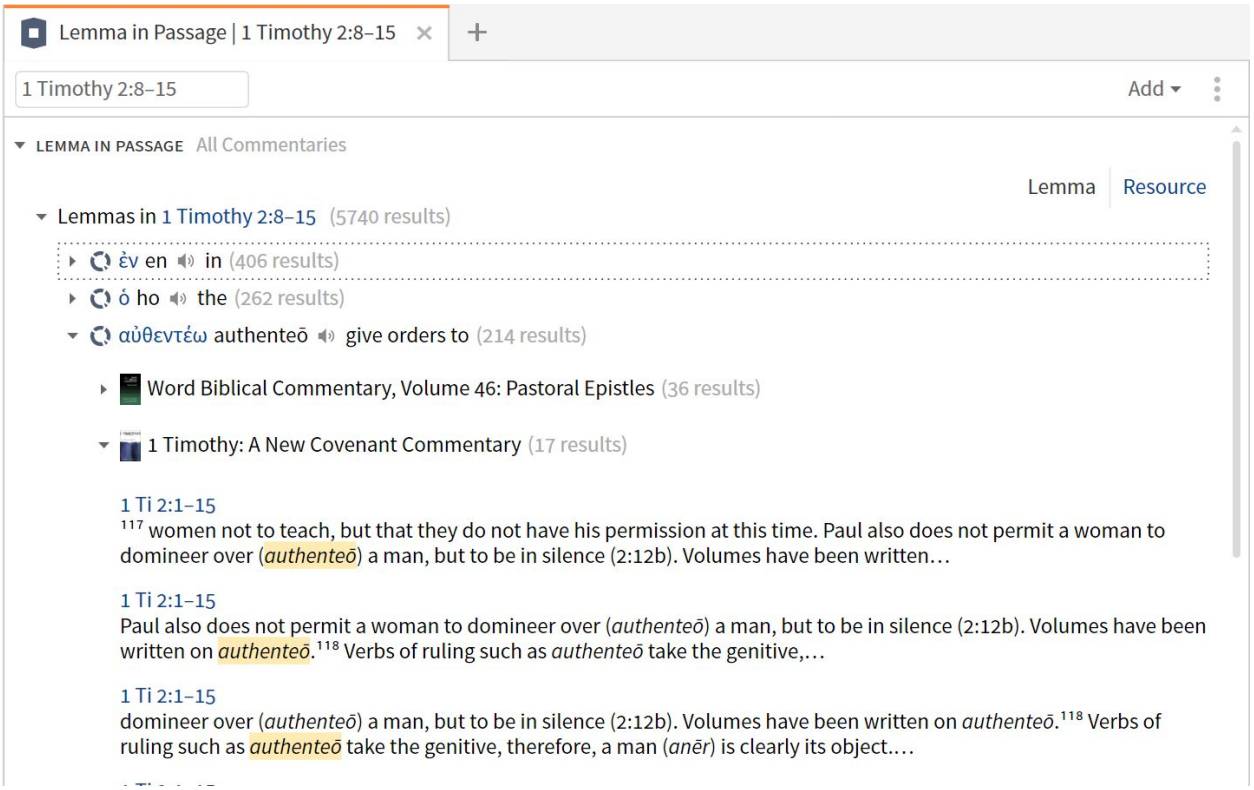
[2] These counts are current as of March 11, 2019.

[3] An oversimplification.

Because we are dealing with commentaries, which are ordered by Bible references, we can remember the Bible reference that the commentary text is associated with. In the Faithlife *lingua franca*, this is known as a "milestone." So if the commentary passage is on John 1:1–4, then that's the milestone; that it the passage that the discussion is anchored to.



This allows us to query the data to determine commentary lemmas for a particular reference range. Since we know the exact location of the string that was lemmatized, we can also easily point the user to the exact commentary context in which the relevant lemmatized word occurs, and the display can pull the location and context of the word for display to users.



And this is really, really great.

But notice the limitations as regards determining terms important to a passage: This points to every commentary word instance for which a lemma can be deduced in the passage. So there are conjunctions, articles, prepositions, and other function words near the top of the list because they are frequently occurring in commentaries that either reproduce the whole text, or commentaries that reproduce the text phrase by phrase.

The Insight

So you've probably already done the math and know where I'm headed with this. These things are usually clearer looking back after the fact than when we're deep in the middle of something.

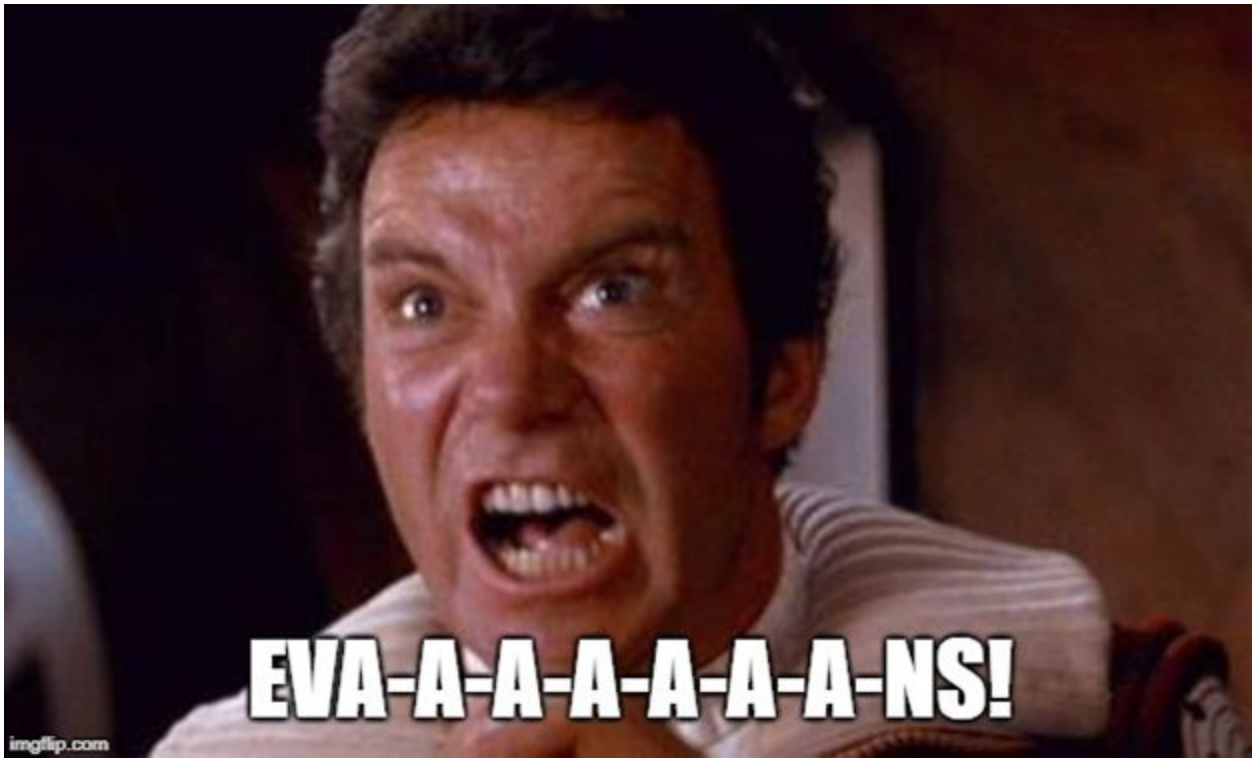
The Lemma in Passage data is a perfect dataset to derive important word data from.

Apparently Eli Evans was the first to arrive at this conclusion after he and I were talking one September day in 2015:

> Eli and I were talking about the "Lemma in Commentary Passage" feature, where we essentially lemmatize original language instances of Greek, Hebrew, and Aramaic words.

> And Eli mentioned that this could be better fodder to serve what we display for Important (sic: “Interesting”) Words.[4]

And Evans was right — he beat me to the punch.



The data positively associates an instance of an original language word *worth discussing* with a passage milestone.

In other words, it gives an authoritative opinion on which words in a given passage are worth discussing because the author of the commentary actually stopped to discuss the word instance. When all these opinions are aggregated across all available commentaries (and Study Bibles), given enough data, one can begin to see which words are typically discussed, which words are infrequently discussed, and which commentaries simply reproduce the original language text — and that even those typically restate terms before discussing them in context.

This information aggregated from all commentaries provides a basis to determine the most important words by using how frequently a word in a passage is discussed as a proxy for importance.

Ok, there’s a little more to it than that (as Peter will show us), but the Lemma in Passage data is the starting point for the data that we eventually used to fuel the Important Words feature.

---

[4] From an internal document from September 2015 that indicates that Eli Evans provided the first flash of inspiration.

The Action

We didn't do much right away. Checking git repos indicates that we first started to look into deriving "Important Word" data from Lemma in Passage data in February of 2017.

We decided to start with the simplest thing that might work, which is usually the best place to start. So we aggregated lemmas per passage and counted frequencies to determine which lemmas were mentioned most frequently in a given passage. We chose to do this across pericopes (we maintain our own set of pericopes to use in situations like this). If the milestone range the lemma occurs in intersects the range of the pericope, then it gets included. Here's an example:


```
{ "pericope-unit": {
  "pericopeReference": "bible+nrsv.75.2.8-75.2.15",
  "pericopeTitle": "Men and Women in the Church",
  "commentaryReferences": {
    "LLS:CBCN1T2TTI": {
      "bible+nrsv.75.2.1-75.2.8": {
        "lemma.lbs.el.ἀνήρ": 3
      },
      "bible+nrsv.75.2.9-75.2.15": {
        "lemma.lbs.el.κοσμέω": 2,
        "lemma.lbs.el.ἡσυχία": 3,
        "lemma.lbs.el.ὑποταγή": 2
      }
    },
    "LLS:29.1.40": {
      "bible+nrsv.75.2.1-75.2.8": {
        "lemma.lbs.el.πᾶς": 15
      },
      "bible+nrsv.75.2.8": {
        "lemma.lbs.el.βούλομαι": 3,
        "lemma.lbs.el.ἀνήρ": 2,
        "lemma.lbs.el.ἐν": 2,
        "lemma.lbs.el.πᾶς": 2,
        "lemma.lbs.el.τόπος": 2,
        "lemma.lbs.el.ἐπαίρω": 4,
        "lemma.lbs.el.χείρ": 4,
```

The list is ordered by commentary (the LLS: identifier), by milestone reference, (the bible+nrsv identifier) and then has a list of lemmas and the number of times they are mentioned in the passage (the lemma.lbs.el identifier).

Frequencies were great, but they didn't do the job. Some of the issue could've been mitigated by using stopwords, but there are times in exegesis where a stopword (preposition or conjunction, perhaps) could actually be meaningful. So it didn't seem prudent to simply rule out particular lemmas globally.

So we tried a TF-IDF variation where each "document" were the lemmas in a pericope, and the "documents" were the set of pericopes across the Bible. This actually did pretty good at providing a score of importance for each lemma, but it wasn't quite good enough.

Also, we had a bunch of other data that could be used to establish word importance — primarily data from lexicons.

Data from Lexicons

With Commentaries, we are retrieving references (lemmas) associated with a milestone (Bible reference). If you think about it, there are other books that are ordered by milestones as well. For our purposes, the next best example to aggregate data from involve original language lexicons.

In this case, we retrieve references (Bible references) associated with a milestone (lemma). This allows us to see which Bible references are frequently used when discussing specific original language words for the task of defining them.

Including these sorts of reference-milestone relationships are important, even though it makes life (where "life" is defined as "creating some sort of importance score based on all of these relationships") difficult.

So we've aggregated all of this data from all of these different resources in a format that can be easily consumed for analysis and, eventually, for scoring Bible words for their importance in particular passages.

The Handoff

How does that happen? This is where we call in the big guns.

The Tech

To put this all together we used Python with Luigi.

https://github.com/spotify/luigi
Luigi is a Python package that helps you build complex pipelines of batch jobs. Created by Spotify, it's basically a Pythonic "Make" utility.

Convergence

Several different sources of knowledge need to converge, and flow together into a single smooth stream of data.

Sources

- Lexicons
- Commentaries
- Theological Words dataset (weight=3)
- Figurative Language dataset (NT only for now)
- Bible Knowledge dataset (Events + People + Places + Things): Counts lemmas that refer to entities participating in an event within a passage describing the event.
- Tf-Idf: How many passages does this lemma occur in? Is it common or rare?

Filtering

Our sources sometimes mention lemmas that are relevant but don't actually occur in the passage being discussed. There's also a little bit of noise caused by difficulty finding the boundaries of relevant text in a commentary. Since our goal is to find the important words that are actually in each passage, we filter out the rest.

From Counts to Scores

So far we've collected raw counts from several sources. How do we turn them into useful scores?

Something like this. Within each passage, add up the scores from all our sources for the words in that passage. But then what?

Step 3

How do we make sure the scores are intuitively comparable, even across different passages? Some passages attract a lot more commentary than others, so they will have higher scores. The least important word in a popular passage could have a higher raw score than the most important word in a less popular passage, which is not what we want.

We also need to support generating a list of important words for an arbitrary passage. Users can select anything from a single verse to an entire book.

Finally, we ought to draw the line somewhere between important and unimportant words. In some sense all the words are important, but listing the "important" words is not very meaningful unless we exclude some.

Normalization

In order to make scores intuitively comparable, we normalize them all to the range from 0-100%.

To deal with the disparity in the amount of data available for different books of the Bible, we normalize the scores on a book-by-book basis. This makes sense because commentaries typically cover a whole book, so the number of commentators covering each passage will be about equal within each book. On the other hand, a book is in most cases long enough that we avoid over-normalizing so much that every passage's most important word has the same score. The most important word in the book is going to have a higher score than the most important word in some other passage in the same book.

We support arbitrary passages, but with some limitations. The smallest unit with comparable scores is a pericope, which can range from a verse to a chapter in length. If a shorter range is requested, we simply filter out any words that don't occur in the requested passage. The longest range we support is an entire book.

(We normalize with logarithmic scale, based on the intuition that if a word is mentioned twice as often as another, that means it's more important, but less than twice as important. I'm not certain this is correct, since the idea of a numerical "importance" is very arbitrary. This affects the scale of the scores, but not the ranking.)

We exclude any words with scores below 50%. There are just a few passages with long long lists of important words, so we cut those off at 40 words.

Math

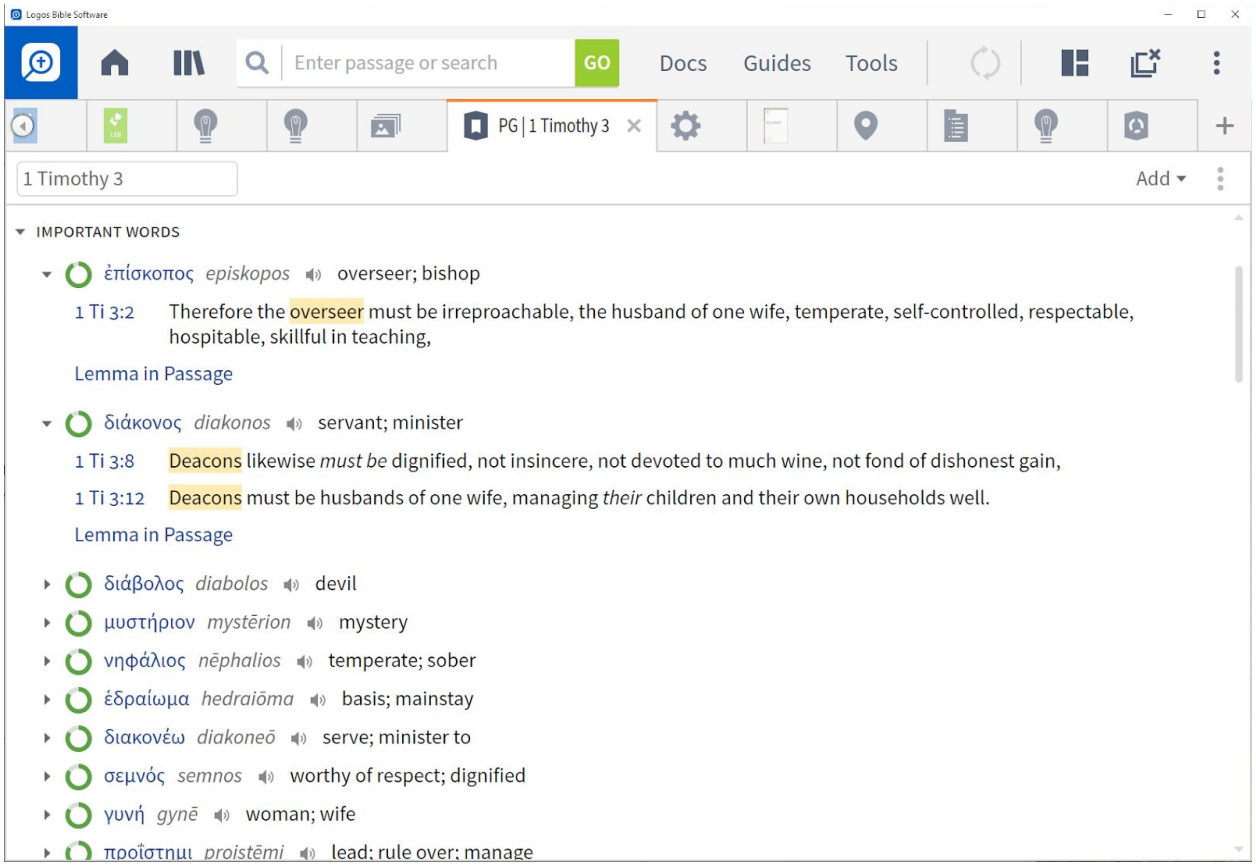
If you're math-averse you can ignore this slide, but for those who like formulas, here it is.

The denominator is the sum over the book, for each book of the Bible.

$$s_w = \frac{log\ c_w}{\sum_c log\ c}$$

The Product

Finally, here's how it looks in Logos 8.


Logos Bible Software
Enter passage or search
GO
Docs
Guides
Tools
PG | 1 Timothy 3
1 Timothy 3
Add
IMPORTANT WORDS
ἐπίσκοπος episkopos overseer; bishop
1 Ti 3:2 Therefore the overseer must be irreproachable, the husband of one wife, temperate, self-controlled, respectable, hospitable, skillful in teaching,
Lemma in Passage
διάκονος diakonos servant; minister
1 Ti 3:8 Deacons likewise must be dignified, not insincere, not devoted to much wine, not fond of dishonest gain,
1 Ti 3:12 Deacons must be husbands of one wife, managing their children and their own households well.
Lemma in Passage
διάβολος diabolos devil
μυστήριον mystērion mystery
νηφάλιος nēphalios temperate; sober
ἑδραίωμα hedraiōma basis; mainstay
διακονέω diakoneō serve; minister to
σεμνός semnos worthy of respect; dignified
γυνή gynē woman; wife
προΐστημι proistēmi lead; rule over; manage