New Testament Hyper-Concordance

Table of contents

The Hyper-Concordance began as a programming exercise to try some corpus linguistics techniques with the Open Scripture Information Standard, a recent XML standard for encoding Bible texts. I wanted to see what i could learn from more directly connecting the words of Scripture together. You can jump right in and explore through the Home Page: this article provides background about what, why, and how.

1. What's a "hyper-concordance"?

A concordance is simply an index of words in a document. Concordances for the Bible were among the first exhaustive indexes ever created for printed texts: John Marbeck published his English concordance in 1550.

The basic idea behind the hyper-concordance is to navigate the space of Scripture directly, using words as links. Most Scripture websites have a search box where you enter a word to find verses that use that word. For example, searching the English Standard Version New Testament for the word "pots" finds two verses, Mark 7:4 and Revelation 2:27 (you need to use the advanced search and select "Exact matches only"). From the standpoint of connecting information, this provides a conceptual link from a single word to one or more verses of Scripture.

Taking this idea one step further, given the text of the verse, you can just embed a hyperlink from the word in question to other verses, preserving the context. Now here's where the idea takes off: instead of just hyperlinking one word, suppose *every* word is hyperlinked? This connects the information and gets you directly from the context of one verse to another with similar content (because of similar words). With some special processing to index the words, every word can link to a list of verses, each word of which is in turn hyperlinked to others, each word of which ... you get the idea.

Here's an example from the page for "Hear" (the links are live into Hyper-Concordance: this example is also on the Home Page):

Mark.4.24	And he said to them, “Pay attention to what you hear: with the measure you use, it will be measured to you, and still more will be added to you.

The unlinked words are function words or other high-frequency terms: they could be linked, but there would be little added value, and it would take a lot more space (the entire hyper-concordance as static HTML amounts to about 80Mb).

Inflected verbs and plural nouns are linked to their base forms (in this example, "measured" -> "measure", "said" -> "say"). Most other Scripture search engines i've seen either match exactly (treating "said" and "say" as two different words), match substrings (for "pot", this

has the peculiar result of matching "spots" and "Mesopotamia"!), or match from the beginning of the word ("pot" matches "pots", "potter", etc.). I wanted to try to do a better job about matching the real dictionary form of words (more about that below).

I've seen this approach used for dictionaries like the HyperDictionary, and it's become a big more common as richer Bible sets are produced. But when i published the first version of this (in May 2003), it was the only example of a thoroughly hyperlinked Scripture concordance that i knew of (email me if you have other information so i can give credit where due).

2. Creating the Hyper-concordance

The hyper-concordance is programmed in Perl using the XML::Twig module for XML parsing of the OSIS sources: you're free to download the source, and it would be easy to adapt to other uses. The input is simply the OSIS version of the ESV text of the New Testament. CSS is used for rendering the pages, so you could create a different look if you wanted to.

The algorithm simply goes through each word of each verse (other than stopwords), and creates a hash table mapping base forms to the verses they occur in. A static HTML page is generated for each term and placed in a directory under the initial letter to keep things more manageable. The same approach is used to generate the index pages, which i find interesting all on their own (but then, i enjoy reading dictionaries too!).

On the index pages that list all the terms starting with a given letter, you can either display the terms using the same size font, or vary the size by frequency, with more frequent terms displayed larger. Note this is *not* the same as websites like del.icio.us or flikr that use variable sizing to indicate the *popularity* of a tag: i have no popularity information at this point, though that would also be a very interesting feature. But this gives you an easy way to find common words, and also provides a preview of how many references you'll find for a given word. If you hover over the link, there's also a tooltip that tells you exactly how many references there are.

3. Linguistic Issues

As indicated above, i wanted to map inflections and plurals back to their dictionary forms. As a practicing computational linguist, it was tempting to use language processing smarts to figure this out, and that's still one way to go. But there's one significant advantage of the New Testament as a subject of corpus analysis: the vocabulary is fixed, and, by most modern standards, remarkably small. I was surprised to find the 8k verses of the New Testament amount to only about 175k words, with a vocabulary size of less than 7k (the exact numbers depend on how you count, of course). Here's the entire vocabulary list (including inflected

verbs and plurals) with counts: as is typical of most natural language problems, vocabulary items aren't normally distributed, but follow Zipf's Law, a topic for another day. By comparison, the University of Pennsylvania Treebank, the foundation of most recent work on trained parsers, contains 1M words of Wall Street Journal text, and these days corpora of 10s of millions of words aren't unusual. With a vocabulary this small, it's entirely feasible to review the whole list in a reasonable amount of time.

So i elected to simply create a list mapping inflected forms and plurals to their bases. I probably missed a few, but it only took a couple of hours. Of course, i'd have to re-do this work for another translation, which would make a more principed approach more attractive. But i doubt any morphological parser for English would be able to tell me that "besought" is the past tense of "beseech"!

Another problem with this approach is that it's hard to know how far to go. Plurals are easy, though i didn't map those without a corresponding singular (like "pangs"). I wasn't completely consistent with words that could be either derived nouns or -ing verbs: should "saying" go back to "say", or stand on its own? And in a few genuinely ambiguous cases, i wound up conflating things i would have rather kept separate: is "lives" the plural of "life", or the inflected form of "live"? It requires more serious processing and understanding context to get these correct. In general i preferred to group things rather than leave them separate, unless there was good reason to do otherwise. I also decided to stop short of mapping superlatives like "better" and "best" back to "good", "happiness" back to "happy", and so forth.

4. Conclusions

Is this useful? Well, i hope so, but i'm not sure (at least i find it *interesting*). This is not a general Bible search tool, and i have no plans to make it one. It doesn't cover the whole Bible (and would be unwieldy for some terms if it did), doesn't let you restrict search to specific sections, etc. But i'm already thinking of ways to extend this. For example, you could view the links between verses as defining a graph, with the strength of relationship determined by how many words are linked in each, and how frequently these words occur in the corpus. It might make an interesting link diagram, though i don't have the software tools to generate one. Another possibility is to add links to each term for other terms with "close" semantics: a topic for another day, though the combination of Nave's Topical Bible and statistical learning techniques might provide for some interesting experiments.

Having created this, i'm very interested to know what people think about it. Please email me with feedback, errors, or constructive criticism.

This story was originally published on the Blogos weblog.