Locating New Testament Cross-References

Some Strategies

Rick Brannan
Logos Bible Software[1]

BibleTech:2008
Bible Technologies Conference
January 25, 2008

Version: 2008-01-23

ABSTRACT

This talk examines the feasibility of locating related passages in the New Testament using various measures. The focus will be on strategy and results, not on the nitty-gritty details of the code.

[1] Author email: **rick logos com**, with **@** and **.** substituted for the spaces, respectively.

INTRODUCTION

Marginal cross-references have long been a feature of several Bibles in print. Each of the myriad versions has some edition with "marginal cross-references" or "center-column cross-references". Yet electronic editions, apart from those reproducing data available in printed editions,[2] have not done a good job of complementing the text with relevant cross-references. Most electronic editions of Bibles are centered on the words of the text and not its presentation or on supplying ancillary data to help in the study of the text.

This paper largely restricts itself to discussing New Testament cross-references the New Testament. Different approaches, from "no-tech" to "low-tech" to (keeping the rhythm) "mo'-tech", will be examined (each in differing degrees). Discussion of necessary data and even ideas about sources are provided at relevant points.

But first, it is necessary to note that there are several different types of cross-references, and perhaps even several different "levels" of cross-referencing. Cross-referencing can be between key words in a text (perhaps even down to key words in a book/author); it can be between similar phrases; it can be topically oriented. But even tables of Gospel parallels are cross-references of a sort.

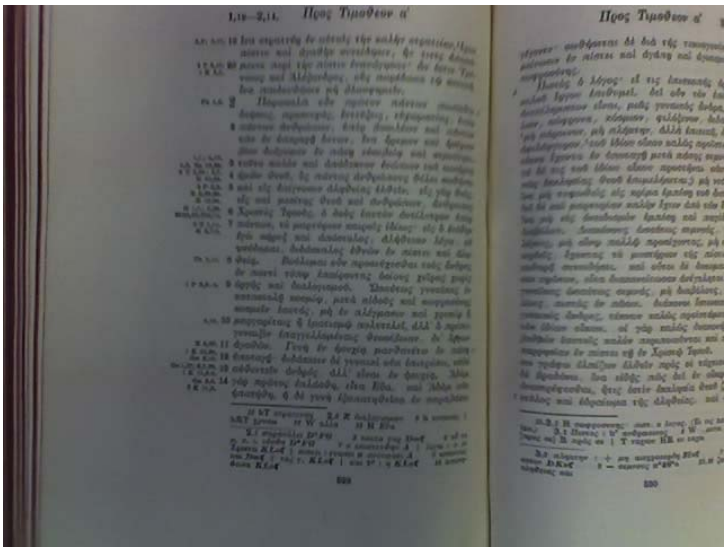
This paper takes a sort of "shotgun" approach, mentioning several ideas on different styles or sources of cross-references and even providing worked examples of many. But we will move quickly from idea to idea. In other words, the presentation will be wide, not deep.

THE NO-TECH APPROACH

This is perhaps the easiest approach, and for many is likely the best approach. Simply put, the approach is "steal it".

Ok, that's a little strong and not quite an accurate description of the approach, but that's the basic idea. Someone else has done the work already. Bibles already have cross-references in them. All that is left to do is to find a decent source and use it (given proper permission). Alternately, one could locate sources that don't require permissions and use them. This approach has the benefit of relying on data that has already been reviewed or edited to some degree. For example, the 1912 printing of Nestlé's Greek New Testament (ninth edition):

[2] Such as the editions provided by Logos Bible Software, which include cross-references supplied by publishers attached to verse or word.



The cross-references (and the base passage) would have to be recorded in some manner, but the data is there and available. Of course, this is the least geeky approach, and arguably the least entertaining. Since we're at a "BibleTech" conference, we'll quickly move along to the fun stuff.

THE LOW-TECH APPROACH

The "low-tech" approaches typically involve reshuffling and processing of available data sources. These can be at the "section" (multiple verse) level, the verse level, the term level, or even at a thematic level.

Section-to-Section Cross References

Gospel Parallels

One example of publicly available cross-references are the Eusebian Canon tables,[3] found in the prefatory material of many editions Greek New Testaments.[4]

What are the canon tables? Simply put, they are the earliest form of a harmony of the gospels that we have. They go back to at least the third century, with roots in the second century:

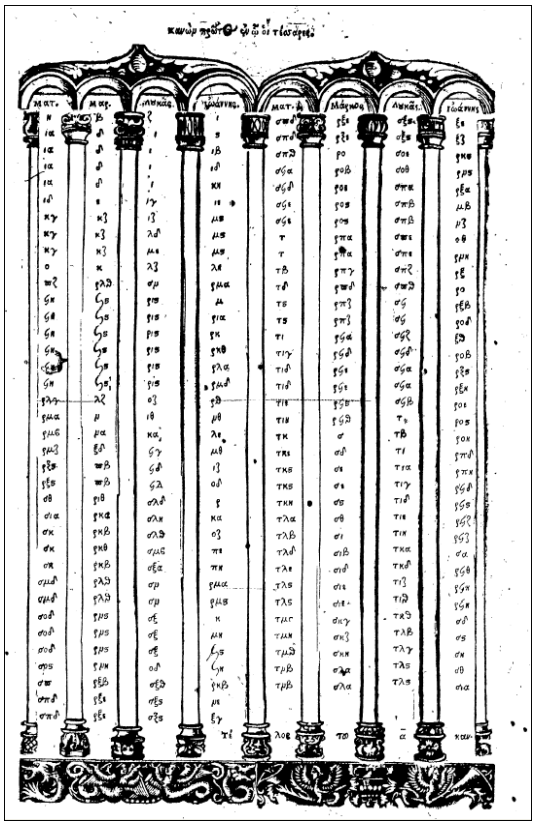
> The system had its roots in the work of one Ammonius of Alexandria, who some time in the second century arranged a sort of partial gospel harmony, taking the text of Matthew as his base and paralleling it with sections of the other gospels. Each section was numbered, and the numbers are referred to as the Ammonian Sections. (Confusingly, the Ammonian Sections are sometimes referred to as *kefalaia*. This usage is to be avoided. Not only is it confusing, but the Ammonian Sections average much shorter than the *kefalaia* — e.g. in Matthew there are 355 sections but only 68 *kefalaia*.)

[3] Alternately known Ammonian Sections; Eusebius, in his letter to Carpatius that describes the canon tables, notes that he got them from Ammonius.

[4] Another example is Sean Boisen's *Composite Gospel Index*. Online: http://www.semanticbible.com/cgi/cgi-overview.html. Accessed January 7, 2008.

> Roughly a century later, Eusebius of Cæsarea (the famous church historian) hit on a scheme to dramatically improve the Ammonian apparatus, by allowing any section of any gospel to serve as the basis point while still letting the reader look up parallels. Starting from the Ammonian divisions (which he may have modified somewhat), he created a set of lookup tables (to use a modern computer term) for finding cross-references. To each Ammonian number, he affixed a canon table number, showing the table in which the reader was to look for the cross-references.[5]

These are typically prefaced with an edition of a letter from Eusebius to Carpatius that briefly explains the system. Editions of the Nestle text have included this information since at least the ninth edition (published in 1912, the earliest I can verify as I happen to own a copy). But the canon tables have been included in print editions for centuries, and in manuscripts for centuries before that. The earliest printed edition of the Greek New Testament, known as the Complutensian Polyglot,[6] contains the letter from Eusebius but does not have the canon tables.[7] Erasmus' edition of 1522, however, does contain the tables. Here is a sample:



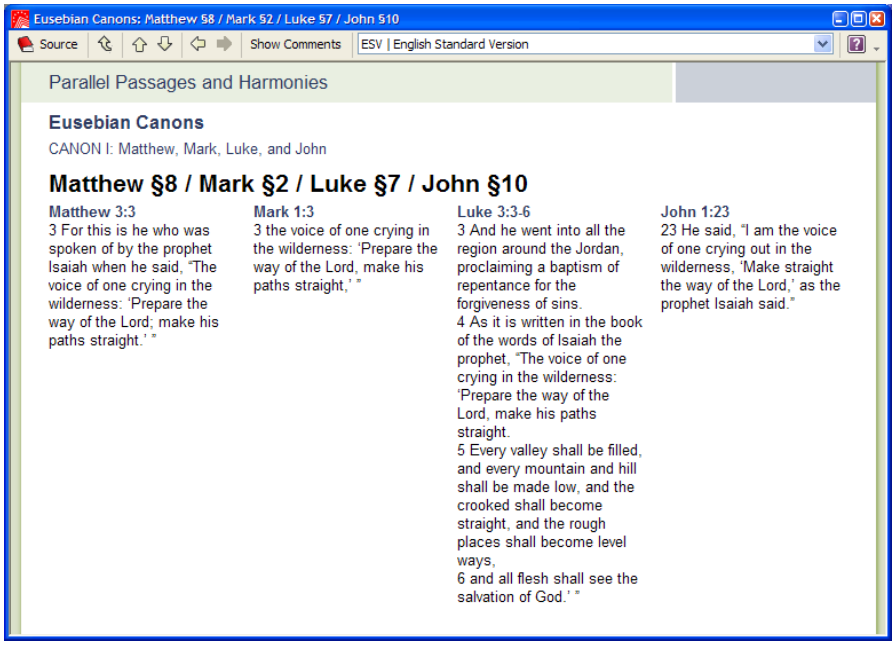
[5] *Divisions of the Text*. Online: http://www.skypoint.com/members/waltzmn/Divisions.html. Accessed Dec. 28, 2007.

[6] While the Complutensian Polyglot was printed first (1514?); the first edition available was that of Erasmus, published in 1516.

[7] As far as I can tell, that is. I have access to an image-only PDF facsimile edition of the Complutensian Polyglot; this facsimile only has the letter in the NT prefatory matter, no tables.

Processing the Data

Where the Eusebian canons appear in print, the numbers representing canon boundaries typically occur in the margin of the running text of the New Testament. So one could follow the margins and create reference ranges for each canon reference. But that's a lot of work. Thankfully, this task has already been done by Kevin P. Edgecomb and he has made the data available on his web site.[8] This allows one to write a program to read the canon table data, define pericope boundaries and associations, and convert the data into something more usable in an electronic environment. Awhile back, here at Logos we did just this.[9] Here is how the underlying data is currently represented in the system:



When a user is researching any one of the verses in question (using the "Passage Guide" feature), the parallel is made known to the user, essentially acting as a cross-reference. These could be converted to cross-references in section headings; they could alternately be converted to marginal cross references with some notation that they represent gospel parallels.

Old Testament Quotations in the New Testament

In the above example, Mt 3.3 is cited. This verse contains a citation of Isaiah 40.3. When Old Testament text is quoted, cited or alluded to in the New Testament, knowledge of the source citation is handy (to say the least). Listing these citations as cross references makes sense as the source context and phrasing may be exegetically valuable.

---

[8] Edgecomb, Kevin P. *The Eusebian Canons*. Online: http://www.bombaxo.com/euspage.html. Accessed Dec. 28, 2007. If you're planning on using this data in an application of some sort (commercial or otherwise) please do contact Mr. Edgecomb for further details.

[9] Though as I recall, we used a different source for the canon data—Tischendorf's *Editio Octavo Critica Maior*, which contains canon references inline in the Greek New Testament text.

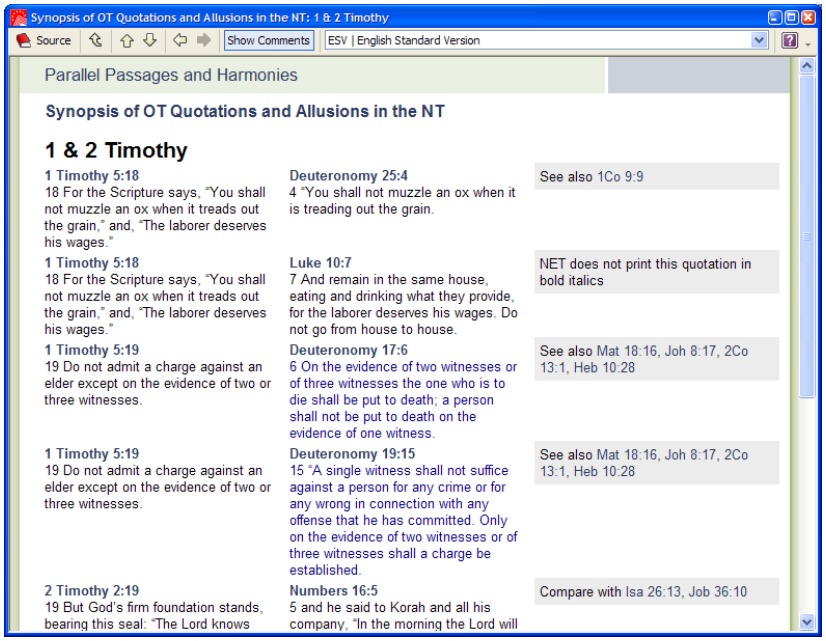
Processing the Data

Data would be processed similar to the method previously descrived for the gospel parallels. Some available online sources include:[10]

- **Blue Letter Bible: Parallel Passages in the New Testament quoted from Old Testament.** http://www.blueletterbible.org/study/misc/quotes.html[11]
- **Michael D. Marlowe. Quotations from and Allusions to the Old Testament in the New Testament.** http://www.bible-researcher.com/quote01.html
- **Joel Kalvesmaki. Table of Old Testament quotes in the New Testament, in English translation.** http://www.kalvesmaki.com/LXX/NTChart.htm

Of course, other resources are available in print. There are too many to mention (see the back matter of virtually any study Bible), but one common and well-accepted source is the UBS4 Greek New Testament, which has a table of quotations and allusions as appendix material. Editions of Westcott & Hort's Greek New Testament include a similar index in the appendix. Most New Testament translations also note quotation source in footnotes.

At Logos, we have several sources of data for Old Testament Quotations and/or Allusions in the New Testament. The primary method of accessing these is through the Passage Guide, as mentioned when discussing canon tables.



[10] Though of course, before integrating such data into commercial applications, proper permission to use data should be obtained.

[11] These are cited as originating in material for the Online Bible.

Thematic Cross References

Some resources, such as *Nave's Topical Bible* (Nave's) and *Torrey's New Topical Textbook* (Torrey's) are available for this purpose. The original editions of both texts are in the public domain; either could serve as a source of data for topical cross-references. Nave's has a huge amount of topics with relatively long lists of references while Torrey's is more brief (630+ top-level topics, 23,000+ sub-topics, more than 21,000 unique references by my calculation); to my mind this makes Torrey's perhaps the better choice to serve as a basis for topical cross-reference data for a given passage.

Processing the Data

Versions of *Torrey's New Topical Textbook* have been available online for years. The Christian Classics Ethereal Library (CCEL, http://www.ccel.org) has an edition in many flavors, notably in plain-text and also their own XML dialect, ThML.[12]

One easy strategy would be to process each group of references (containing more than one reference) as a cross-reference group, implementing these in the same manner as marginal cross references. For example, the topic "Adoption: Is through Christ" has references of Jn 1.12; Ga 4.4-5; Eph 1.5 and Heb 2.10, 13.

```
56
57      Adoption
58          Explained -- 2Co 6:18.
59
60          Is according to promise -- Ro 9:8; Ga 3:29.
61
62          Is by faith -- Ga 3:7,26.
63
64          Is of God's grace -- Eze 16:3-6; Ro 4:16,17; Eph 1:5,6,11.
65
66          Is through Christ -- Joh 1:12; Ga 4:4,5; Eph 1:5; Heb 2:10,13.
67
68          Saints predestinated to -- Ro 8:29; Eph 1:5,11.
```

This simple approach would, for a reference, insert the balance of the references as cross-references for the group. So Jn 1.12 would have cross-references of Ga 4.4-5; Eph 1.5 and Heb 2.10, 13.[13]

Term-to-Term Cross References

While thematic cross references have to do with concepts discussed within a section (typically a verse) of Bible text, some cross references are more granular, linked at the term level. These sorts of potential cross references are concordance-style in nature, but are not simply listings of same-spelled words. They have some association by sense, or have been further disambiguated (e.g. multiple people sharing the same name).

---

[12] *Torrey's New Topical Textbook | Christian Classics Ethereal Library*. online: http://www.ccel.org/ccel/torrey/ttt.html. Accessed December 28, 2007. Note that in recent years CCEL has changed its usage policy from free to anyone for any use to a non-commerical use license; so if one plans to use data originating at CCEL in a commercial application, permission should be obtained prior to making public releases.

[13] Small note of historical interest: At one point in time (early 2000s?) I actually had a web site (now defunct) called FreeBibleResources.com that did just this sort of thing. For a given reference it provided a list of topics and further references on each topic, using Torrey's as source. I called it the "Study Starter". The website died, but the concept lives on in Logos Bible Software's "Home Page" and its "Bible Study Starter" feature.

Term Reference by Louw-Nida Semantic Domain

In 1988, the United Bible Society published Johannes Louw & Eugene Nida's *Greek-English Lexicon of the New Testament Based On Semantic Domains*. (henceforth LN or Louw-Nida).[14] Most lexicons provide further information on words; the Louw-Nida lexicon provides further information on semantic sense. Thus, instead of being organized alphabetically, the lexicon is organized semantically, with 93 top-level 'domains' ranging from "Geographical Objects and Features" to "Names of Persons and Places". The order of domains is from wide to specific, and ordering within domains follows the same pattern. Each domain is numbered (from 1-93) with many domains further divided into lettered sub-domains. Each article within each domain is further numbered. Each entry thus has a unique two-part numeric identifier of *domain.article*. Many words have more than one entry because many words have more than one semantic sense. For example, γενναω (give birth) has four entries—'beget' at 23.58, 'give birth' at 23.52, 'be born of' at 13.56, and 'cause to happen' at 13.129. The phrase γενναω ανωθεν (born again) has its own entry at 41.53.

If each word in the Greek New Testament is tagged with its LN identifier, words could be concorded by their semantic sense. This could provide interesting fodder for cross-references, particularly for infrequently-used senses or other items of significance.

Processing the Data

In my own long-term study of the Pastoral Epistles, I did exactly this sort of annotation. I only considered verbs, nouns and adjectives within the Pastoral Epistles and tagged them with what I considered to be the proper LN identifier. I built this data into a concordance, which is available online.[15]

---

[14] Louw & Nida, *Greek-English Lexicon of the New Testament Based on Semantic Domains*. New York: United Bible Societies. 1988.

[15] Brannan, Rick. *A Concordance of the Pastoral Epistles ordered by Semantic Domain*. Online: http://www.supakoo.com/rick/pastorals/indexes/louwconc.htm. Accessed January 4, 2008.

6. Artifacts

B. Instruments Used in Agriculture and Husbandry (6.4–6.9)

LN 6.8

1 Ti 6:1 (word #4): ζυγὸν (*zygon*)

E. Traps, Snares (6.23–6.25)

LN 6.23

1 Ti 3:7 (word #16): παγίδα (*pagida*)
1 Ti 6:9 (word #9): παγίδα (*pagida*)
2 Ti 2:26 (word #7): παγίδος (*pagidos*)

J. Instruments Used in Marking and Writing (6.54–6.67)

LN 6.59

2 Ti 4:13 (word #16): μεμβράνας (*membranas*)

LN 6.118

2 Ti 2:20 (word #8): σκεύη (*skeuē*)
2 Ti 2:21 (word #9): σκεῦος (*skeuos*)

Of course, this approach has the same problem that other concordance-style approaches have. For terms (here semantic senses) that occur frequently, some distinction must be made as to the most relevant references to include as including the whole list is not feasible.

Person Names and Place Names

These are typically people and places (ranging from buildings and streets to cities and larger geographic regions like countries and even mountain ranges or bodies of water). Internally at Logos we have databases of people and places that are being used in all sorts of ways; these datasets are forming the basis of something we're calling the Bible Knowledgebase. [16] Sean Boisen's New Testament Names database is available online;[17] OpenBible.info has a database of Bible places available as well.[18] Alternately; Louw-Nida domain 93 is all about People (subdomain A) and Places (subdomain B); so they have already identified and provided at least some references for all people and places in the New Testament.

Processing the Data

Processing the data would be very similar to other approaches documented previously. With people and place names, however, comes the problem of common or popular people and places. If one's dataset has disambiguated people and places, then this is not a problem. If starting with raw concordance data,

[16] The "Biblical People" feature in Logos Bible Software is fueled by an instance of our Biblical people database; this plus other sources are being used by Sean Boisen to form the Bible Knowledgebase. For more info, hear a paper by Sean at this conference; alternately head to the Bible Knoweldgebase section of his own web site: http://semanticbible.com/blogos/category/bible-knowledgebase/

[17] Boisen, Sean. *New Testament Names: A Semantic Knowledge Base*. Online: http://www.semanticbible.com/ntn/ntn-overview.html. Accessed January 7, 2008.

[18] *Bible Geocoding*. Online: http://www.openbible.info/geo/. Accessed January 7, 2008.

however, disambiguation is necessary. Whatever dataset is used, some work is necessary to prune large occurrence lists (say, greater than 10 instances) down to the most salient or significant instances.

Low-Tech Conclusion

There are several available resources, particularly at the section-to-section, term-to-term and thematic levels. These, with minimal additional processing and perhaps some editorial work, can become the basis of a larger set of NT cross references.

THE MO'-TECH APPROACH

The "Mo' tech" approach involves computing term significance or similarity between textual groups (verses, pericopes, etc.). In all cases, the text used as basis of comparison is the UBS[4] edition of the Greek New Testament. Where morphology and lexical forms are required, the Logos morphology is used.

Word-to-Word Cross References

Statistically Improbable Words

In any given text, some words occur more frequently; other words occur less frequently. There are statistical measures that can help in evaluating these words. This approach compares the use of a word in a given book of the New Testament with its usage across the whole NT. It highlights words that have an improbable frequency in a given book (so they occur more frequently than expected), and words that have an improbable infrequency in a given book (so they occur less frequently than expected). These words whether improbably frequent or improbably infrequent, may have an important role to play in the book—thus knowing other occurrences of the same word in the book could be valuable exegetically.

This approach uses the z-score, primarily because it is easy to explain (thus easy to understand) and relatively easy to calculate. There are other statistical measures that may be more appropriate to apply to this problem, such as the log-likelihood test.[19]

Processing the Data

James Tauber, responding to a question posted to the B-Greek email list regarding vocabulary acquisition approaches in September of 2006, helpfully explains what a z-score is and how to calculate it.[20] The relevant portion of Tauber's post, with equations and explanation, is excerpted below:

```
z = (p_hat - p_0) / sqrt ((p_0 * (1 - p_0)) / n_t)

where
```

[19] See Ted Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," Computational Linguistics 19(1), 1993. Online: http://citeseer.ist.psu.edu/rd/0%2C29096%2C1%2C0.25%2CDownload/http://citeseer.ist.psu.edu/cache/papers/cs/7119/http:zSzzSztina.lancs.ac.ukzSzucrelzSzpaperszSztedstats.pdf/dunning93accurate.pdf. Accessed January 7, 2008

[20] Tauber, James. [B-Greek]tool for vocabulary distribution of GNT books? Online: http://lists.ibiblio.org/pipermail/b-greek/2006-September/040322.html. Accessed December 28, 2007.

```
p_hat = t / n_t  [ratio of word in subtext]
p_0 = o / n_o    [ratio of word in overall text]

where

t = occurrences of word in subtext
o = occurrences of word in overall text
n_t = total number of words in subtext
n_o = total number of words in overall text

results: : 0 would mean as expected
           positive means more frequent than expected
           negative means less frequent than expected.
```

So, using a z-score, one can statistically determine words that occur with improbable frequency (high positive score), and with improbable infrequency (low negative score).

Using Tauber's roadmap, I wrote some code to build a list of z-scores for each lexical form in the New Testament. Requisite counts were computed, variables in Tauber's equations filled in, and a list of z-scores for each lexical form in the NT, sorted by occurrence in each book, was generated. Consider the below, which has data for the word λογος.

```
16557      </lemma>
16558      <lemma l="λογομαχία">
16559          <book num="75" name="1 Ti" cnt="1" cntbk="1591" z="9.20666133021866" />
16560      </lemma>
16561      <lemma l="λόγος">
16562          <book num="61" name="Mt" cnt="33" cntbk="18346" z="-1.64237905568155" />
16563          <book num="62" name="Mk" cnt="24" cntbk="11304" z="-0.583023108310189" />
16564          <book num="63" name="Lk" cnt="32" cntbk="19482" z="-2.13891578434403" />
16565          <book num="64" name="Jn" cnt="40" cntbk="15635" z="0.428598595032643" />
16566          <book num="65" name="Ac" cnt="65" cntbk="18450" z="3.14853233810087" />
16567          <book num="66" name="Ro" cnt="7" cntbk="7111" z="-2.42862105494834" />
16568          <book num="67" name="1 Co" cnt="17" cntbk="6830" z="0.165936304955998" />
16569          <book num="68" name="2 Co" cnt="9" cntbk="4477" z="-0.52154404559323" />
16570          <book num="69" name="Ga" cnt="2" cntbk="2230" z="-1.4446579556323" />
16571          <book num="70" name="Eph" cnt="4" cntbk="2422" z="-0.745106048046095" />
16572          <book num="71" name="Php" cnt="4" cntbk="1629" z="5.33333299490217E-02" />
16573          <book num="72" name="Col" cnt="7" cntbk="1582" z="1.65634120060193" />
16574          <book num="73" name="1 Th" cnt="9" cntbk="1481" z="2.90448232562662" />
16575          <book num="74" name="2 Th" cnt="5" cntbk="823" z="2.16420179097444" />
16576          <book num="75" name="1 Ti" cnt="8" cntbk="1591" z="2.1539360401679" />
16577          <book num="76" name="2 Ti" cnt="7" cntbk="1238" z="2.35100889807876" />
16578          <book num="77" name="Tit" cnt="5" cntbk="659" z="2.73130630744765" />
16579          <book num="79" name="Heb" cnt="12" cntbk="4953" z="0.04584763388253" />
16580          <book num="80" name="Jas" cnt="5" cntbk="1742" z="0.409610427852257" />
16581          <book num="81" name="1 Pe" cnt="6" cntbk="1684" z="0.98475430882048" />
16582          <book num="82" name="2 Pe" cnt="4" cntbk="1099" z="0.847609641155195" />
16583          <book num="83" name="1 Jn" cnt="6" cntbk="2141" z="0.389835472811653" />
16584          <book num="85" name="3 Jn" cnt="1" cntbk="219" z="0.659121578368272" />
16585          <book num="87" name="Re" cnt="18" cntbk="9851" z="-1.14563647302861" />
16586      </lemma>
16587      <lemma l="λόγχη">
16588          <book num="64" name="Jn" cnt="1" cntbk="15635" z="2.63457041851017" />
16589      </lemma>
16590      <lemma l="λοιδορέω">
16591          <book num="64" name="Jn" cnt="1" cntbk="15635" z="0.812434270621533" />
```

In the above, the *cnt* attribute is the count of the lexical form in the book; the *cntbook* attribute represents the total number of words in the book. The score to note is the 5.33 z-score of Philippians. In this book, λογος is used more frequently than expected. Romans, on the other hand, with a -2.4286 z-score, shows that λογος occurs less frequently than expected. Philippians has the most statistically

improbable usage, though its usage is statistically improbable because of its frequency. Romans has the most improbable infrequent usage.

From here, book-specific reference lists of statistically improbable words can be generated. The list could be further refined by only including severe outliers (say, words with a z-score of abs(5) or greater) and filtering out words by part-of-speech (say, only include verbs, nouns, adjectives and adverbs). These lists could serve as basis of term-specific cross references designed to highlight and bring to prominence when particular words are used in improbable frequencies; this may have some value for study.

```
653         <lexeme lemma="λογίζομαι" inflection="λογιζόμεθα" z="11.8013028651359" />
654         <lexeme lemma="γάρ" inflection="γὰρ" z="12.385965288298" />
655         <lexeme lemma="δικαιόω" inflection="δικαιοῦσθαι" z="9.16570717483127" />
656         <lexeme lemma="πίστις" inflection="πίστει" z="7.77331372384931" />
657         <lexeme lemma="νόμος" inflection="νόμου" z="20.2592311592404" />
658     </verse>
659     <verse dt="bible.66.3.29" ref="Ro 3.29">
660         <lexeme lemma="θεός" inflection="θεὸς" z="10.3862755411366" />
661         <lexeme lemma="καί" inflection="καὶ" z="-9.19347271837016" />
662         <lexeme lemma="ἔθνος" inflection="ἐθνῶν" z="7.15314687892398" />
663         <lexeme lemma="καί" inflection="καὶ" z="-9.19347271837016" />
664         <lexeme lemma="ἔθνος" inflection="ἐθνῶν" z="7.15314687892398" />
665     </verse>
666     <verse dt="bible.66.3.30" ref="Ro 3.30">
667         <lexeme lemma="εἴπερ" inflection="εἴπερ" z="4.8398549631638" />
668         <lexeme lemma="θεός" inflection="θεὸς" z="10.3862755411366" />
669         <lexeme lemma="δικαιόω" inflection="δικαιώσει" z="9.16570717483127" />
670         <lexeme lemma="περιτομή" inflection="περιτομὴν" z="9.65337334985557" />
671         <lexeme lemma="πίστις" inflection="πίστεως" z="7.77331372384931" />
672         <lexeme lemma="καί" inflection="καὶ" z="-9.19347271837016" />
673         <lexeme lemma="ἀκροβυστία" inflection="ἀκροβυστίαν" z="9.82196887481834" />
674         <lexeme lemma="διά" inflection="διὰ" z="9.68457852185708" />
675         <lexeme lemma="πίστις" inflection="πίστεως" z="7.77331372384931" />
676     </verse>
677     <verse dt="bible.66.3.31" ref="Ro 3.31">
678         <lexeme lemma="νόμος" inflection="νόμον" z="20.2592311592404" />
679         <lexeme lemma="οὖν" inflection="οὖν" z="4.40419783450924" />
680         <lexeme lemma="καταργέω" inflection="καταργοῦμεν" z="3.9080971291462" />
681         <lexeme lemma="διά" inflection="διὰ" z="9.68457852185708" />
682         <lexeme lemma="πίστις" inflection="πίστεως" z="7.77331372384931" />
683         <lexeme lemma="μή" inflection="μὴ" z="3.60507095521267" />
684         <lexeme lemma="ἀλλά" inflection="ἀλλὰ" z="6.31626913933011" />
685         <lexeme lemma="νόμος" inflection="νόμον" z="20.2592311592404" />
686     </verse>
687     <verse dt="bible.66.4.1" ref="Ro 4.1">
```

It should be noted that using z-scores to identify statistically improbable words should be the start of the process, not the end.

Phrase-to-Phrase Cross References

N-grams and Repeated Word Groups

Many times, cross references involve lexical "hooks"—similar phrases or keywords. One method to examine, therefore, is repetition of similar phrases. This method examines each verse (as defined by the NA/UBS text) and builds N-grams of each verse. An N-gram is basically a group of N words that occur in succession.[21] Consider John 1.1:

> In the beginning was the word, and the word was with God, and the word was God

[21] A more technical definition and discussion of n-grams is available at Wikipedia: http://en.wikipedia.org/wiki/Ngram.

Where N = 5; the verse has the following N-grams (or in this case, 5-grams):

```
In the beginning was the
the beginning was the word
beginning was the word and
was the word and the
the word and the word
word and the word was
and the word was with
the word was with God
word was with God and
was with God and the
with God and the word
God and the word was
and the word was God
```

Processing the Data

These sorts of lists are compiled for the whole of the New Testament, for each verse. The N-grams are then compared to each other and where they match a cross-reference is generated. Instead of acting on English text, the underlying Greek text is used as basis. In addition, instead of the inflected form of the Greek word, the dictionary form is used,[22] making the matching a bit more flexible.

One further step involves canonicalizing the tokens: shifting the words to lower-case and sorting them. This allows tokens to match even when word order is different between tokens. An additional (experimental) step is to, for one of the words in the sequence, include part-of-speech notation instead of the dictionary form; permuting the possibilities. So, instead of "In the beginning was the" we could instead have "PREP the beginning was the" or, when sorted case-sensitively, "PREP-beginning-the-the-was".

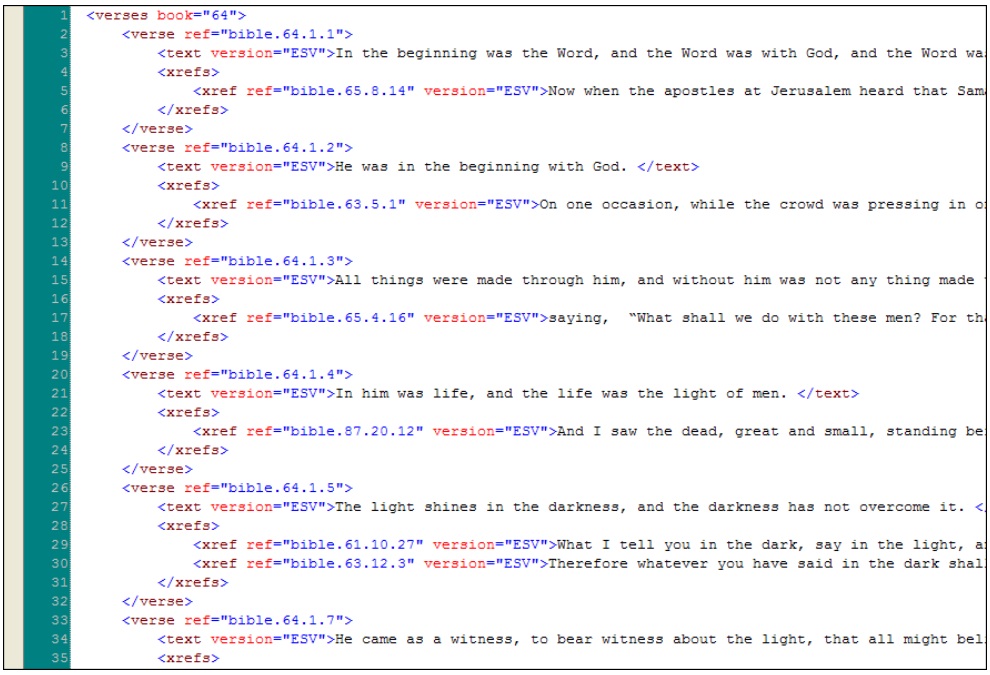
Admittedly, the signal-to-noise ratio is somewhat high with this process (in comparison to starting with existing human-edited cross-reference lists); but the results aren't all bad.

> **Reference:**
> Jn 1.9: The true light, which enlightens everyone, was coming into the world.
>
> **Cross-References:**
> Jn 1.4: In him was life, and the life was the light of men.
> Jn 8.12: Again Jesus spoke to them, saying, "I am the light of the world. Whoever follows me will not walk in darkness, but will have the light of life."

The whole NT has been processed using this method; one of the resulting XML files includes the ESV translation for the verses in question:

---

[22] I've used data proprietary to Logos, known internally as the Logos Greek New Testament Morphology; however one can access a complete morphological annotation of the Greek New Testament at James Tauber and Ulrik Petersen's site, MorphGNT (http://morphgnt.org). The data at the site is available for non-commercial use; if one desires commercial application of the data then one should contact the site owners for permission. Alternately, morphology from Dr. Maurice Robinson for several editions of the Greek New Testament is available at http://users.mstar2.net/broman/editions.html. Note, however, that Dr. Robinson's editions do not use dictionary (lexcial) forms and instead use Strongs numbers; so they may require more processing to be useful. Alternate (and updated) text and morphology information from Dr. Robinson for his 2005 edition of the Majority Textform is available at http://www.rpbyztxt.com.



```
<verses book="64">
    <verse ref="bible.64.1.1">
        <text version="ESV">In the beginning was the Word, and the Word was with God, and the Word wa
        <xrefs>
            <xref ref="bible.65.8.14" version="ESV">Now when the apostles at Jerusalem heard that Sam
        </xrefs>
    </verse>
    <verse ref="bible.64.1.2">
        <text version="ESV">He was in the beginning with God. </text>
        <xrefs>
            <xref ref="bible.63.5.1" version="ESV">On one occasion, while the crowd was pressing in o
        </xrefs>
    </verse>
    <verse ref="bible.64.1.3">
        <text version="ESV">All things were made through him, and without him was not any thing made
        <xrefs>
            <xref ref="bible.65.4.16" version="ESV">saying,  "What shall we do with these men? For th
        </xrefs>
    </verse>
    <verse ref="bible.64.1.4">
        <text version="ESV">In him was life, and the life was the light of men. </text>
        <xrefs>
            <xref ref="bible.87.20.12" version="ESV">And I saw the dead, great and small, standing be
        </xrefs>
    </verse>
    <verse ref="bible.64.1.5">
        <text version="ESV">The light shines in the darkness, and the darkness has not overcome it. <
        <xrefs>
            <xref ref="bible.61.10.27" version="ESV">What I tell you in the dark, say in the light, a
            <xref ref="bible.63.12.3" version="ESV">Therefore whatever you have said in the dark shal
        </xrefs>
    </verse>
    <verse ref="bible.64.1.7">
        <text version="ESV">He came as a witness, to bear witness about the light, that all might bel
        <xrefs>
```

These sorts of references could form the basis of a new set of cross-references, with the final product having undergone some sort of human editing to sift out the chaff from the wheat.

Common Substrings

This is a variation to the N-gram appraoch, comparing each verse to every other verse in the New Testament, keeping track of the number of characters held in common between the two verses.[23] Verses that have many characters in common (usually phrases) are considered related. This is similar to but not exactly the same as the N-gram approach mentioned above.

Processing the Data

Within each verse, dictionary forms (aka "lemmas" or "lexical forms") of words were used, not inflections. Additionally, only verbs, nouns, adjectives and adverbs were considered in each verse. Accents and breathing marks were removed, and all words were lower-cased.

The similarity was calculated using the following approach:

let $a = length of first verse (number of characters)
let $b = length of second verse (number of characters)
let $c = sum of the length of common substrings[24]

$sim = ($c * 2) / ($a + $b)

Verses with a similarity of 50% or greater[25] are considered cross-references. An example of this is found in 1Ti 2.4:

---

[23] Because of time constraints for the paper, I was only able to compare First and Second Timothy as well as separately compare Matthew to Mark as tests to see how viable this approach would be.

[24] Alternately, the length of the longest common substring could be used.

> who desires all people to be saved and to come to the knowledge of the truth. (1 Ti 2:4, ESV)

This is similar to both 2Ti 3.7 and 2.25:

> always learning and never able to arrive at a knowledge of the truth. (2 Ti 3:7, ESV)

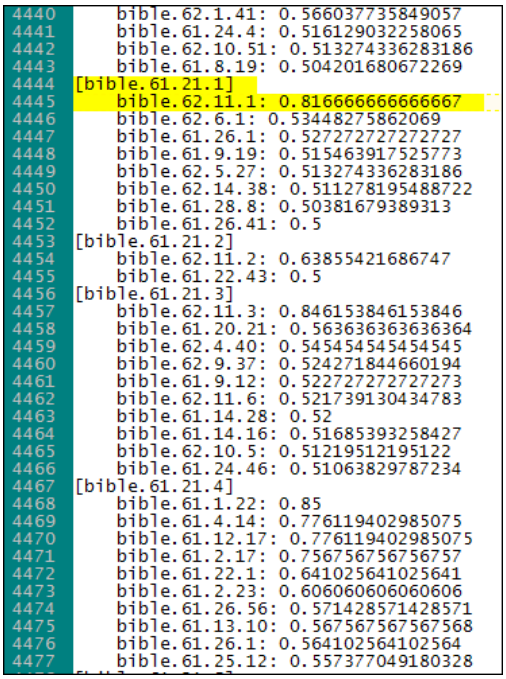
> correcting his opponents with gentleness. God may perhaps grant them repentance leading to a knowledge of the truth, (2 Ti 2:25, ESV)

The longest common substring for all of these verses is ἐπίγνωσιν ἀληθείας ἐλθεῖν, translated "to come into knowledge of the truth". This commonality causes the common substring approach to consider the verses releated, hence their inclusion as cross-references. Here are some more examples; note that `bible.75` is First Timothy; `bible.76` is Second Timothy:

```
[bible.75.2.1]
    bible.75.5.5: 0.521739130434783
[bible.75.2.2]
[bible.75.2.3]
    bible.76.4.7: 0.518518518518518
[bible.75.2.4]
    bible.76.3.7: 0.679245283018868
    bible.76.2.25: 0.5
[bible.75.2.5]
    bible.76.3.17: 0.568807339449541
    bible.75.4.10: 0.507692307692308
    bible.75.1.14: 0.504504504504504
    bible.75.3.13: 0.5
    bible.76.1.13: 0.5
    bible.75.1.2: 0.5
[bible.75.2.6]
[bible.75.2.7]
    bible.76.1.11: 0.605504587155963
[bible.75.2.8]
[bible.75.2.9]
[bible.75.2.10]
[bible.75.2.11]
[bible.75.2.12]
[bible.75.2.13]
[bible.75.2.14]
[bible.75.2.15]
    bible.76.1.13: 0.543859649122807
[bible.75.3.1]
    bible.75.4.9: 0.512820512820513
    bible.76.2.11: 0.506329113924051
```

Alternately, a longest-common-substring comparison of Matthew to Mark locates some synoptic parallels between the two, among other things. In the below excerpt, `bible.61` is Matthew, `bible.62` is Mark. Below note the alignment between Matthew 21 and Mark 11; the similarity between Mt 21.1-3 and Mk 11.1-3; but Mt 21.4's lack of similarity with anything at all in Mark. This hints that similarity data could be used (and likely has been used somewhere) to locate synoptic parallels. Groups of verses with strong similarities between them could indicate parallel passages; this information is helpful and therefore good for cross-referencing purposes.

[25] This number is somewhat arbitrary and could be revised upon further consideration of a more complete data set.



Also interesting is the similar verbiage used in Mt 1.22 and 21.4. As this approach simply looks for similarity by substring, this sort of repetition is found as well. Here are the verses in question.

> All this took place to fulfill what the Lord had spoken by the prophet: (Mt 1:22, ESV)
>
> This took place to fulfill what was spoken by the prophet, saying, (Mt 21:4, ESV)

Such cross-references could come in handy when studying one passage and then desiring to find other locations where prophecy was fulfilled with the fulfilment described in similar language.

This method does equate length with significance, which can be a problem. Some verses can be associated by common terminology, theme, or name; this method will not locate such items.

Using Three-Word Phrases

This approach is another variation on the N-gram and similar substring approaches. Instead of searching out the longest common substring between two verses; it instead compiles all three-adjacent-word combinations (tri-grams), building a concordance of three-word phrases. A few years back I did some initial work in this area,[26] centered on authorship and style issues and not directly on issues of cross-referencing, looking at three-word phrases in the Pastoral Epistles and in the "genuine" Pauline epistles. But the same sort of method could be used to generate potential cross-references using three-word phrases as the point of commonality.

[26] A blog post describing this work, from Dec. 2004, is online: http://www.supakoo.com/rick/ricoblog/2004/12/05/WeekendProjectConcordanceOfThreeWordPhrasesInThePastoralEpistles.aspx. Further posts are online; search the blog for the word "tri-log".

In the data for the Pastoral Epistles, there are 3270 potenial combinations of three-word phrases.Of these, 141 occur more than once.[27] To give a further idea of scope, the so-called "genuine" Paulines (Romans-Second Thessalonians and Philemon) contain 27,422 three-word phrases; 2436 of these (so, less than 10%) occur more than once. Of those 2436, 280 are found in both the Pastoral Epistles and the balance of the Paulines.[28] So this approach would not be a major source of cross-references, but they could provide meaningful cross-references for given portions of text.

Processing the Data

I did initial processing of subsets of data (Rom-2Th; Phm and the Pastoral Epistles) in 2004. The concordances I created at that time provide a glimpse of the sort of data that could be used for cross-referencing purposes. Below is a sample.

αἰών ὁ αἰών

Occurs 5x.

Inflected Phrase: ... αἰῶνας τῶν αἰώνων...	
Galatians 1:5	ᾧ ἡ δόξα εἰς τοὺς αἰῶνας τῶν αἰώνων ἀμήν
Inflected Phrase: ... αἰῶνος τῶν αἰώνων...	
Ephesians 3:21	αὐτῷ ἡ δόξα ἐν τῇ ἐκκλησίᾳ καὶ ἐν Χριστῷ Ἰησοῦ εἰς πάσας τὰς γενεὰς τοῦ αἰῶνος τῶν αἰώνων ἀμήν
Inflected Phrase: ... αἰῶνας τῶν αἰώνων...	
Philippians 4:20	τῷ δὲ θεῷ καὶ πατρὶ ἡμῶν ἡ δόξα εἰς τοὺς αἰῶνας τῶν αἰώνων ἀμήν
Inflected Phrase: ... αἰῶνας τῶν αἰώνων...	
1 Timothy 1:17	τῷ δὲ βασιλεῖ τῶν αἰώνων ἀφθάρτῳ ἀοράτῳ μόνῳ θεῷ τιμὴ καὶ δόξα εἰς τοὺς αἰῶνας τῶν αἰώνων ἀμήν
Inflected Phrase: ... αἰῶνας τῶν αἰώνων...	
2 Timothy 4:18	ῥύσεταί με ὁ κύριος ἀπὸ παντὸς ἔργου πονηροῦ καὶ σώσει εἰς τὴν βασιλείαν αὐτοῦ τὴν ἐπουράνιον· ᾧ ἡ δόξα εἰς τοὺς αἰῶνας τῶν αἰώνων ἀμήν

Comparing to the N-gram approach

The example of 1Ti 2.4, "into knowledge of the truth" was used in discussing the N-gram approach above. What sort of data does the three-word-phrase (or "tri-gram") approach come up with for this same example? Recall the modified N-gram approach located three related passages: 1Ti2.4; 2Ti 2.25; 3.7. Using the three-word-phrase approach, 1Ti 2.4 is contained in two different listings as there are two different trigrams: εἰς ἐπίγνωσις ἀλήθεια (1Ti2.4; 2Ti 2.25; 3.7) and ἐπίγνωσις ἀλήθεια ἔρχομαι (1Ti2.4; 2Ti 3.7).

While the three-word-phrase approach finds the same references, it does so across two listings. Why? Because of the smaller granularity, the inclusion of all lexemes, and the restriction to in-text order (recall the N-gram approach actually sorted a subset of lexical items within the N-gram before comparing). Each approach will contain entries and associations that the other approach misses; the super-set of both

---

[27] R.W. Brannan, *A Concordance of Tri-Logs in the Pastoral Epistles*. Online: http://www.supakoo.com/rick/pastorals/indexes/PastoralsPhraseIndex.htm. Accessed January 4, 2008.

[28] R.W. Brannan, *A Concordance of Tri-Logs Held in Common between the "Genuine" Paulines and the Pastoral Epistles*. Online: http://www.supakoo.com/rick/pastorals/indexes/PhraseIndexCompared.htm. Accessed January 4, 2008.

approaches could however be a decent set of references to further sift and edit and use as source for a new set of NT cross-references.

Using Modifiers in OpenText.org SAGNT

Programmatically delimiting the extent of a phrase or clause is somewhat unreliable, but when a corpus has been annotated with such boundaries, relying on clause and phrase boundaries is possible. In the *OpenText.org Syntactically Analyzed Greek New Testament*, the phrase boundary work has essentially been done in the word group annotation level. Even better, modification relationships within each word group have also been annotated. Grouping not only where the same word(s) occur, but where they occur in similar modification relationship to each other holds more promise than the typical adjacent-word N-gram approach for similar size groups of words.

Processing the Data

Here is a breakdown of how the word μεσιτης (mediator) is modified, according to the *OpenText.org SAGNT*:

This first pass has only associated a word (in this case, μεσιτης) with the words that further modify it. Further work could be done to directly associate a term with the word it directly modifies; these words, as collocates, could be statistically analyzed (perhaps using the log-liklihood ratio test) to determine which modification relationships are significant. The significant instances, then, could be considered cross-references.

In the above example, it is notable that διαθηκη directly modifies μεσιτης in both Heb 8.6 (“covenant he mediates”) and 9.15 (“mediator of a new covenant”). Simply preparing a concordance of modifying collocates could bring such instances to light and could also be used as cross-reference source material. Additional analysis to determine statistically significant modifying collocates could provide further insight.

Section-to-Section Cross References

In the area of section-to-section cross-references, one method that may have promise is the use of a vector similarity algorithm to establish similarity between 'documents' (here sections/pericopes). Unfortunately, I did not have time to explore this option in depth so I cannot report on it here.

The approach does hold promise, though. The idea is to use document comparison measures to determine how similar documents (in this case pericopes) are; results would then be grouped by similarity. This should, in theory, cause the most similar pericopes to group together; and the relationships between these different sections/pericopes could be interesting. For example, one would expect gospel pericopes that are similar to group together.

Mo'-Tech Conclusion

There are several interesting possibilities; the above are only a glimpse of approaches and variations of approach to the problem of locating referential points within a common text.

Generating cross references from the text itself is possible though it does introduce some static due to its nature. Such generated data may be useful as a starting point for compiling cross-references, though further human editing—both deleting extraneous references and perhaps adding references that are appropriate—would be necessary.

CONCLUSION

Hopefully this paper has shown that there are several sources to consult and exploit when it comes to considering cross-references for the New Testament. The ideas mentioned in this paper are only the tip of the iceberg; there are surely a myriad of possibilities, both in the creation and analysis of potential datasets.

In the creation of a cross-reference database, however, one item has become clear to me as I've considered the different approaches in this paper. Cross-references are helpful, but recording the reason for the cross-reference can also be just as helpful. How many times have you followed a cross-reference only to come across a passage that doesn't seem to have much of anything to do with the referring passage? Recording the reason for the cross-reference—be it because of similar words, similar phrasing, common names or places, common themes—is something that automated methods can begin to accomplish. As new and updated cross-reference databases are created and enhanced for the next century of Bible study, schemes for communicating the *what* (reference) along with the *why* (reason) should be part of the equation.