A New Numbering System for Greek New Testament Lexemes

Draft: Id: new-numbering-system.tex 380 2006-03-28 12:28:29Z jtauber

James Tauber
soon to be at
Department of Linguistics
University of Essex
jtauber@jtauber.com

Ulrik Petersen
Department of Communication and Psychology
University of Aalborg
Kroghstræde 3, 9220 Aalborg East, Denmark
ulrikp@hum.aau.dk

Abstract

Numbering systems (such as Strong's) are a popular way to reference the lexemes of the Greek New Testament corpus but a straight enumeration is not without problems, particularly when there is disagreement about whether two forms are the same lexeme or not. We present a way of referencing lexemes that allows competing viewpoints to be represented simultaneously. Existing numbering systems can be mapped into this new system without any loss of granularity and new analyses can be expressed without violating the integrity of existing references into the system.

1 Introduction

In the late 19th century, a concordance of the King James Bible was produced under the direction of James Strong (Strong, 1890). This concordance provided a comprehensive cross-reference of every word of the King James text back to the corresponding words in the original texts. A dictionary was then included that provided a glossary for the various lemmas in Hebrew, Aramaic and Greek. The cross-referencing was achieved by assigning each lemma a number.

While there has been considerable criticism made of the dictionary itself over the last century, there is no doubt that the numbering system he devised for that dictionary has proved useful independent of the definitions themselves.

The mapping of numbers to lexemes provides a means of referencing lexemes in a way that is unambiguous with regard to homographs (distinct words of identical spelling). It also keeps users of the system isolated from different choices of lemma—for example, how to handle deponency (Taylor, 2004) or whether to use the first person present active indicative of a verb or, say, the aorist infinitive.[1] Decisions like this can be made without affecting the integrity of analyses that refer to lexemes by number.

Limitations of Strong's numbers, however, have long been recognized and there have been both attempts to improve Strong's numbering and to develop alternative numbering systems (Goodrick and Kohlenberger, 1990). These limitations (and hence the nature of their correction) generally amount to errors of omission or disagreement over where to draw the line between certain lexemes (see Section 2 for examples).

However, in all cases, the improvements or alternatives remain simple enumerations of a set of lexemes. Any decision as to whether to distinguish two lexemes or conflate them into one is locked into the numbering system itself. A disagreement in how to slice the lexicon up results in competing numbering systems which prevents external databases and analyses based on one to be integrated with those based on another.

What is needed is a manner by which alternative viewpoints can be represented at the same time in a single data structure that can additionally be integrated with existing num-

[1] While the traditional approach has been to identify verb lexemes by the former, there have been strong arguments made in favour of the latter (Buth, 2004; Taylor, 2004).

bering systems. Furthermore, such a system must allow that future analyses which introduce new distinctions or conflations be accommodated without invalidating either the earlier datastructure or references made to it.

2 Some Motivating Examples

Before presenting our solution, let us review a number of examples from the Greek New Testament corpus which highlight the types of issues a new numbering system must deal with.

Consider the words ἀδιαφθορία, ἀφθονία and ἀφθορία which are textual variants of one another in Titus 2.7. They are also all *hapax legomena*, appearing no where in the GNT but that verse. G/K (Goodrick and Kohlenberger, 1990) assign the numbers 91, 916, 917 respectively to these lemma. Strong's, based as it is on the *Textus Receptus*, lists only ἀδιαφθορία and assigns it 90. Any text containing the variant ἀφθονία or ἀφθορία (such as NA27) cannot reference Strong's unless, as some have done, they reuse 90, thus giving it a new meaning of "any of ἀδιαφθορία, ἀφθονία or ἀφθορία."

It is not the case that G/K is always finer-grained than Strong's, though. The adjective βασίλειος meaning 'royal' can be used as a neuter noun βασίλειον meaning 'palace'. Are these two different words? Strong's makes a distinction (934 versus 933) whereas G/K does not (assigning 994 to both). These so-called *cross-over adjectives* are fairly common and neither Strong's nor G/K is entirely consistent in their handling of them. There are some that G/K distinguishes that Strong's does not (see below).

Thus it is seen that differences in numbering systems arise not only because of dealing with different sets of data but also because of differences of interpretation in what constitutes a distinct lexeme.

Even a simple example of μέχρι(ς) can lead to different interpretations. Are μέχρι and μέχρις distinct lexemes or the same? Strong's says they are the same (3360) whereas G/K distinguishes them (3588, 3589). And yet, when it comes to ἄχρι(ς), G/K assigns just one number (948).

Strong's distinguishes ζηλωτής (2207) from Ζηλωτής (2208) but G/K assigns them both 2421.

Σολομών and Σολομῶν differ only in accentuation in their nominative but their genitives differ. Morphological analyses would want to treat these as distinct but, in G/K, they are assigned the same number (5048).

Spelling differences such as ἀνάπειρος versus ἀνάπηρος matter in some applications (e.g. textual criticism, morphology) but not others (e.g. lexical semantics). Both Strong's and G/K treat these two as the same (376 and 401 respectively).

βάτος is a homograph with two senses: a liquid measure and a type of bush. Both Strong's and G/K give these senses distinct numbers. However, the second sense is masculine or feminine depending on dialect and no attempt is made in Strong's or G/K to distinguish this dialectal difference.

These are just a handful of examples demonstrating the kinds of problems we hope to solve with our system.

3 Data Structure

The key to our solution is that each number refers to either a word or a set of other numbers. It is important that no external distinction is made between these two types of numbers.[2]

Because numbers can refer to sets of other numbers and because this relationship is acyclic, it can be viewed as a partial-ordering or lattice.

Because the structure of this lattice is more important than the particular numbers used, we will often use "node" and "number" interchangeably.

3.1 Numbering

We begin by assigning numbers to words as shown in figure 1.

Figure 1: Numbering words

If we stopped there, we may have produced a slight improvement over Strong's or G/K but

[2]This property is key to introducing finer-grained distinctions to the numbering system at a later date (see Section 3.3).

our system would suffer from the same problems.

3.2 Joining

Consider the case where A and B are treated by some as being the same word. With our lattice approach, we create a new node that references the set $\{278, 1326\}$ as shown in figure 2

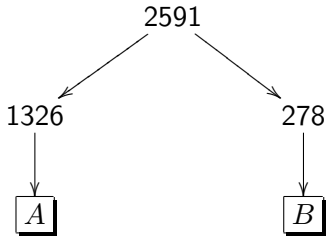

Figure 2: Joining nodes

Note that A and B can still be referred to individually using 1326 and 278. However, if no distinction is desired, 2591 can be used.

This takes care of cases where a coarser-grained reference is required. But what about where certain analyses requires a more fine-grained reference?

3.3 Splitting

In the case where a finer-grained distinction is required, new nodes for each distinct object can be created with the parent node remaining for when a reference makes no distinction. In figure 3, the word assigned 278 is further refined into 2592 and 2593. The removal of a distinct word B doesn't break any external references as the only thing referencing it directly was the node 278. Anything referencing 278 will continue to work with the desired interpretation of $\{B_1, B_2\}$.

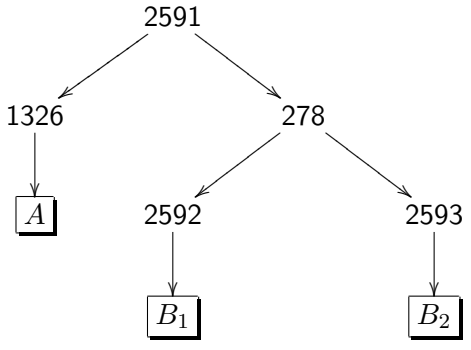

Figure 3: Splitting nodes

Note that, for this to work, the number itself must not give any hint as to whether it is a reference to a word or a set of other numbers.

3.4 Multiple Parents

Although we have not yet found an example which requires it between Strong's and G/K, there is nothing in our approach which limits a node to having only one parent.

In figure 3 we have the ability to reference the conflation of A, B_1 and B_2 as 2591. But what if we wanted to treat A and B_1 together but distinct from B_2? We can achieve this by adding a join node between 1326 and 2592 as shown in Figure 4.

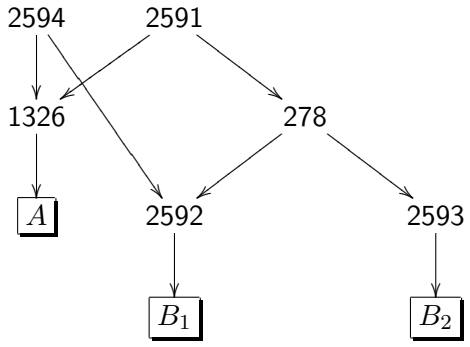

Figure 4: Multiple parents

We see no reason to disallow this kind of flexibility.

4 Applying the Approach to the Greek New Testament

Prior to our collaboration, the authors had both worked independently on data integration tasks with the Greek New Testament corpus and various analyses. Tauber predominantly used G/K numbering for this purpose while Petersen used Strong's.

During the course of this independent work, we had already augmented our respective lexicons to include alternative lemmas and so we had a starting point for building a lexeme lattice that incorporated the differing viewpoints of both Strong's and G/K.

However, no attempt had been made to distinguish whether two forms with the same number were alternative lemmas or whether they were distinct lexemes that Strong's or G/K had conflated. So the first task in applying the lexeme lattice approach was to disambiguate these cases.

Figure 5 shows an extract of the augmented G/K listing before the merge took place.

1680:ἐκγαμίσκω
1681/1:ἔκγονον
1681/2:ἔκγονος
1682:ἐκδαπανάομαι
1682:ἐκδαπανάω
1683:ἐκδέχομαι

Figure 5: Augmented G/K Listing Before Merge

The two lines marked 1682 are an example of alternative lemmas for what is clearly the one lexeme. 1681, on the other hand, is an example where G/K had conflated what might be considered distinct lexemes by some and so one was (temporarily) labelled 1681/1 and the other 1681/2.

This file and the corresponding augmented Strong's list were then merged using a script written in the Python programming language that attempted to find corresponding entries for the same lexeme, based either on an identical match of lemmas or the entry in one list being a subset of the other (with the missing lemmas not appearing in any other entry).

The result is show in figure 6.

ἐκγαμίσκω:1548:1680
ἔκγονον:None:1681/1
ἔκγονος:1549:1681/2
ἐκδαπανάομαι|ἐκδαπανάω:1550:1682
ἐκδέχομαι:1551:1683

Figure 6: Merged Strong's and G/K

Fields are delimited by colons with the first column the lemma (or lemmas, further delimited by vertical bars), the second column the Strong's number (or occasionally numbers, delimited by vertical bars) and the third column the G/K number (or numbers).

This file still requires considerable manual correction. Even in the extract in Figure 6 it can be seen that ἔκγονον and ἔκγονος should probably be given Strong's of 1549/2 and 1549/1 respectively, given that presumably Strong's has simply conflated the two under the single lemma ἔκγονος.

The corrected, merged, file will then becomes the basis for the lexeme lattice. We are still in the process of making corrections and, once this is done, we will publish a preliminary version of the lattice. The plan is to also provide a lookup service on our website.[3]

5 Below the Lexeme

So far we have not yet considered one other significant way in which numbering systems can differ in their analyses. Strong's numbering system, in particular, will often assign different numbers to forms of the same lexeme if those forms are distinct enough.

Fortunately, the lattice structure described in Section 3 is equally suitable for handling differing treatments of stem alternation and irregular forms as it is for handling differing ideas of what the boundaries between lexemes are.

Consider the words εἷς and μία. These are, respectively, the masculine and feminine forms of the numeral 'one'. By most accounts, these are the same lexeme, but Strong's assigns them distinct numbers. One approach would be to say that this distinction is not necessary and μία can be viewed as just an alternative lemma for the εἷς lexeme. In this case both Strong's 1520 and 3391 would be mapped to the same node in our lattice. However, it is equally possible in our system to give εἷς and μία each their own node and then provide a join node that can be used when no distinction is intended.

An even more interesting case, is that of the personal pronouns. ANLEX (Friberg, Friberg and Miller, 2000) and G/K treat first person ἐγώ as distinct from second person σύ. Strong's goes a step further and has separate entries for different cases and number, spelling out almost the entire paradigm.

The nodes, then, need not stop at the finest-grained notion of a lexeme. The broad strokes of the lexeme's paradigm can be sketched out at the next level or, if need be, the entire paradigm can be included in the structure.

Including particular forms as children of the lexeme node itself provides a way of representing principal parts. It also provides a way to handle suppletion.

φάγω and ἐσθίω can be given distinct nodes, with a join node being used for the lexeme and,

[3] see `http://morphgnt.org/`

where desired, the suppletive stems available for referencing.

Other examples include λέγω and its second aorist form εἶπον, distinguished by Strong's but not G/K, and ὁράω and its aorist form εἶδον, distinguished by G/K but not Strong's.

6 Conclusion

Any attempt to integrate independent analyses of a corpus *post hoc* is likely to be confronted with incompatible distinctions being made. This is clearly the case with lexemes in the Greek New Testament. While consensus in what constitutes a distinct lexeme is desirable or even required in much linguistic work, there is tremendous value in providing a means to augment this with competing analyses so that independent work can be more easily integrated.

We believe the approach we have described is a significant step forward in allowing more precise referencing of lexical entries that can also be integrated with other numbering systems regardless of differering levels of granularity or disagreements over where to draw the boundaries between lexemes.

The lattice-based numbering system proposed truly allows one to "have their cake and eat it too" by enabling coarse-grained and fine-grained divisions to be represented in the same structure.

Furthermore, the system is robust to future analyses which may seek to make further incompatible distinctions. Indeed, such *post hoc* distinctions need no longer be viewed as incompatible and can be made without affecting the integrity of references to the prior lattice.

We hope to soon be publishing an initial lattice for the lexemes of the Greek New Testament, with mappings from both the Strong's and G/K numbering systems.

The new numbering system will then become the basis for our future lexical work and hopefully that of others.

References

James Strong. 1890. *Strong's Exhaustive Concordance of the Bible Strong's Exhaustive Concordance of the Bible With Hebrew Chaldee and Greek Dictionaries*

Edward W. Goodrick and John R. Kohlenberger. 1990. *The NIV Exhaustive Concordance* Zondervan, Grand Rapids, MI.

Randall Buth. 2004. *Verbs Perception and Aspect: Greek Lexicography and Grammar.* in Bernard A. Taylor et al., *Biblical Greek Language and Lexicography* Eerdmans, Grand Rapids, MI.

Bernard A. Taylor. 2004. *Deponency and Greek Lexicography* in Bernard A. Taylor et al., *Biblical Greek Language and Lexicography* Eerdmans, Grand Rapids, MI.

Timothy Friberg, Barbara Friberg, and Neva F. Miller. 2000. *Analytical Lexicon of the Greek New Testament* Baker Books, Grand Rapids, MI.