Automatic Manuscript Comparison Prototype

Andrew Edmondson

Introduction

Inspired by DNS sequences, I decided to explore the idea of encoding NT manuscripts in a similar way: ABCABDCD... etc.

The concept is as follows: Each verse will count as a unit, including the title (verse 0) of a work. Each unique version of that verse then gets assigned a letter, a, b, c etc. Then any text can be encoded and compared quickly with any other.

Texts

Transcriptions of manuscripts are downloaded from http://www.igntp.org/ in XML form. These are then parsed and each verse is indexed and assigned a letter. 07 (Codex Basiliensis) was chosen as a base text, somewhat arbitrarily, as it is Byzantine and therefore likely to be more similar to the majority of other manuscripts than, say, 01 (Codex Sinaiticus).

The base text of the first verses of John 1 is then represented as:

```
MS 07            aaaaaaaaaaa...
```

Comparing 04 (Codex Ephraemi rescriptus) with 07 we see:

```
MS 07            aaaaaaaaaaa...
MS 04              dabaaaaaa...
```

The first two verses of 04 are missing, and verse 3 is just a fragment. It can then be seen that v4 is the same in 07 and 04. In the web-representation of the output, hovering over a letter shows the text of the verse in that manuscript. So, for example, v5, is as follows:

07: και το φως εν τη σκοτια φαινει και η σκοτια αυτο ου κατελαβεν

04: και το φως εν τη σκοτεια φαινει και η σκοτεια αυτο ου κατελαβεν

Correctors as states of the text

The XML transcriptions not only show the final state of the text, but where possible also show the original and details of corrections made by different hands. 01, for example, can be represented like this:

```
MS 01 (firsthand) abbbabbbbbbbcbcbbbbbbbbbbbbbbcbbbbbbbbbbbbbbbbbbbbb
MS 01 (S1)              c          d  d   c     c      c      c
MS 01 (S2)                         e
MS 01 (ca)           a   c    a   d g cc   d     cc   c    c   cd cc ca cd    c
MS 01 (cb2)                        f          a    c                 c
MS 01 (corrector)                                                       c
```

This shows that v3 has been corrected (by "ca") to be the same as 07 (although this, of course, says nothing about a direct relationship between 01 and 07). The corrector changed "ουδε εν" to "ουδεν".

Verse 15 immediately stands out as it has been changed by 4 of the 5 different correctors at work on 01 – and each corrector has changed it to a different text form:

c (firsthand) : ϊωαννης μαρτυρι περι αυτου και κεκραγεν om ουτος ην om ο οπιcω μου ερχομενος οc εμπροcθεν μου γεγονεν οτι πρωτος μου ην

d (S1) : ϊωαννης μαρτυρι περι αυτου και κεκραγεν λεγων ουτος ην ο ειπων ο οπιcω μου ερχομενος εμπροcθεν μου γεγονεν οτι πρωτος μου ην

e (S2) : ϊωαννης μαρτυρι περι αυτου και κεκραγεν ουτος ην ον ειπων ο οπιcω μου ερχομενος εμπροcθεν μου γεγονεν οτι πρωτος μου ην

g (ca) : ϊωαννης μαρτυρι περι αυτου και κεκραγεν ουτος ην ο οπιcω μου ερχομενος om εμπροcθεν μου γεγονεν οτι πρωτος μου ην

f (cb2) : ϊωαννης μαρτυρι περι αυτου και κεκραγεν ουτος ην ον ειπον ο οπιcω μου ερχομενος εμπροcθεν μου γεγονεν οτι πρωτος μου ην

We can then say that this verse warrants more detailed study – why has it been changed so much in this one manuscript? What kind of changes are they? etc.

A bigger picture

The following table shows the comparison of the IGNTP's first 17 majuscule transcriptions:

```
MS 07 (firsthand)    aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
MS 07 (2)                                          a
MS 07 (corrector)                 b b
MS 01 (firsthand)    abbbabbbbbbbcbcbbbbbbbbbbbbbcbbbbbbbbbbbbbbbbbbbbbbb
MS 01 (S1)                 c          d   d   c     c      c      c
MS 01 (S2)                            e
MS 01 (ca)               a   c    a   d g cc   d     cc   c    c   cd cc ca cd    c
MS 01 (cb2)                           f          a    c                 c
MS 01 (corrector)                                                                c
MS 02 (firsthand)    aacaadaacaaaecbaaeccacaccddadaadcdcadecddcdacdeaccd
MS 03 (firsthand)    aaacaecababcfdhbdddcedcddeededcedddbefdeededdefcdde
MS 03 (2)               d          e           b
MS 03 (secunda_manu) a                                                          d
MS 04 (firsthand)      dabaaaaaaadfibefedfeabeffefedfeeacfgef
MS 04 (2)                             j   g egc ac g gfegf    gif
MS 05 (firsthand)    aaebafdcdabdggkc
MS 05 (A)                  c
MS 05 (B)                        eh l
MS 05 (K)              f
MS 09 (firsthand)    b da   ed cafihbaaafahcaeaahaaaahaacadjaaaefeafeafaf
MS 09 (corrector)       e                                                 a
MS 011 (firsthand)   aaaaaaaaaaaadibafhgaicacagigaafgafaddkggafgfafeagef
MS 011 (corrector)                                      g   g                f
MS 013 (firsthand)   aaaacafaad gdjbaaiafafaaaahaaaaaaacadaaafghaefgaaaf
MS 017 (firsthand)   aaaaaagaaeaadfmaeahgacaaaadhhgaiaaaahlhaghiaffhefff
MS 017 (corrector)                                        d i
MS 019 (firsthand)   caaadaaaeaaajkbbejidfcdfdhjadhgjgeeafmjihiaaggifhgg
MS 019 (corrector)                                              ji
```

```
MS 021 (firsthand)   daaaaagaaacadfnaakahagaafadaiahkaeaadjaajejahfjafha
MS 022 (firsthand)                          jhegcikicailheaadnk
MS 024 (firsthand)                                  jagmaacedola
MS 028 (firsthand)   aaafaagaaaaaefmaegaakcaaaaladaadaeaadkiafjdaifkaiaf
MS 028 (corrector)                                              a
MS 029 (firsthand)                               agamjkagn
```

It is quite straightforward to spot Byzantine manuscripts by the amount of “a” entries (Although this actually means that the verse is identical to the text in 07. A different base text might yield different results).

Several questions or points of interest are raised by considering the data above. For example, what was 028's corrector's reason for changing v39 to the “a” text. Why is the “d” text so popular in v37? Or the “c” text in v9? Etc.

Next steps

The next steps in this prototype may include:

1. Further automatic regularisation – e.g. nomina sacra. The lack of such regularisation is currently corrupting the results somewhat, so while I can develop a method of analysis some of the conclusions may be spurious at this stage.
2. Finding a way to represent the degree of difference for each text form of a verse – is it radically different or has just a single letter changed?
3. Finding a way to represent the type of difference – addition, omission, etc.
4. Integrate with the complete set of IGNTP manuscripts – e.g. thus including papyri
5. ?