Rationale and Implementation of the Collation System Used on this CD-ROM

Peter Robinson

1. Beginnings: before 1998

There is not a rich literature concerning text comparison (collation) strategies and systems.[1] This is not because few people ever compare texts: the many thousand scholarly editions which include some record of variants texts show that for centuries scholars have routinely compared one text with another. It is rather, I suspect, that the processes of text comparison seem so trivially routine as not to be worth writing about. Collation seems a simple matter. Just chose a base text and compare it with other versions, noting the differences in each from the base text, and that is all.

In essence, this was our collation system up to 1998. We constructed what we called a 'base text for collation.' We then compared this word by word with each witness. For each witness, we compiled a list of all the differences between the witness and the base text. We gathered all these lists together so that we could identify all the witnesses which agreed with the base at particular points, and also identify all the witnesses which agreed with each other in a particular variant. We used computers to help with the collation, storing the lists of variants, adjusting the regularization of variant spellings so as to create a 'substantive' collation, but we did not question the basis of our collation. Accordingly, the collation on our first CD-ROM, of The Wife of Bath's Prologue published in 1996, followed this model entirely.

Here is an example of the results from this collation system. In line 35 of The Miller's Tale we find the following versions of the line among the 54 witnesses which have it:

This Carpenter hadde wedded newe a wyf	*25 witnesses*
This Carpenter hadde wedded a newe wyf	*23 witnesses*
This Carpenter hadde newe wedded a wyf	*1 witness*
This Carpenter hadde wedded newly a wyf	*2 witnesses*
This Carpenter hadde E wedded newe a wyf	*1 witness*
This Carpenter hadde newli wedded a wyf	*1 witness*
This Carpenter hadde wedded a wyf	*1 witness*

Our procedure was to collate these 54 witnesses each against the base text, listing those which agreed and disagreed with the base, as follows:

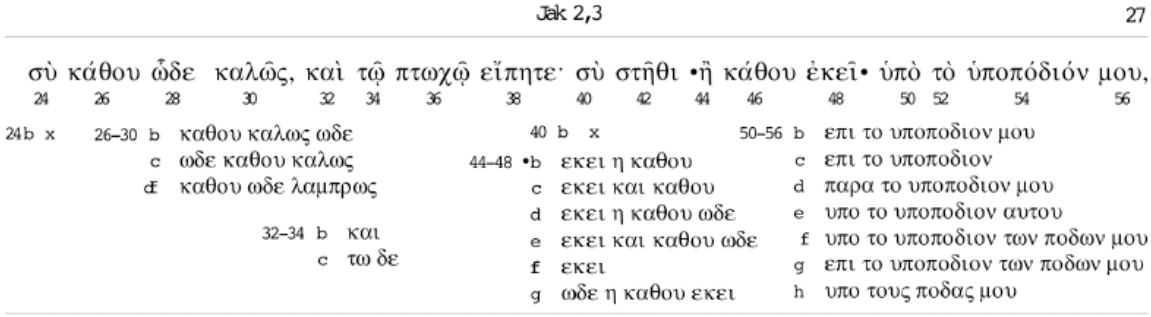
This] *54 witnesses*
Carpenter] *54 witnesses*
hadde] *54 witnesses*
wedded] *53 witnesses*; E wedded *1 witness*
wedded newe] newe wedded *1 witness*, newli wedded *1 witness*
newe] *26 witnesses*; newly *1 witness*; omitted *1 witness*
newe a] a newe *23 witnesses*
a] *30 witnesses*
wyf] *53 witnesses*

This is efficient, and certainly satisfactory in terms of recording variants with reference to the base text. We see here that for the first three words and the last word there is no variation, and we just state accordingly that all witnesses there agree with the base and with each other. All the variation occurs on the three base text words 'wedded newe a'. This variation is actually recorded against five lemmata: in turn 'wedded', 'wedded newe', 'newe', 'newe a' and 'a'. Observe that the phrases 'wedded newe' and 'newe wedded' both overlap one other, and also overlap the three words 'wedded' 'newe' 'a'.

2. Problems with base text collation

In 1998 two events occurred which disturbed my confidence in this collation method. The first was a visit to the Institut

für neutestamentliche Textforschung in Münster, Germany, to examine with them the possibility of applying the methods we had developed for the *Canterbury Tales* to their work on the Greek New Testament. In the course of this analysis, I had the opportunity to study closely the system they use to represent textual variation. At that point, they had just published the first volume of their ambitious (and splendid) *Editio Critica Maior* of the Greek New Testament, drawing together the evidence of 182 manuscripts of the letter of James (Aland 1997). Here is a fragment of their printed collation:



Notice here the treatment of the variants on the phrase κάθον ὧδε καλῶς. Two of the variants represent different word ordering; the third combines a different ordering of two of the three words with a variant on the third. In our system, the variants on the base might appear as follows:

κάθον] b d; ὧδε c
ὧδε] d; καλῶς b; κάθον c
καλῶς] c; ὧδε b; λαμπρως d

By comparison, one could express the same information using the Münster system with all variants registered against the one three-word lemma κάθον ὧδε καλῶς:

κάθον ὧδε καλῶς]
κάθον καλῶς ὧδε b
ὧδε κάθον καλῶς c
καλῶς ὧδε λαμπρως d

It is immediately clear that it is much easier for the reader to see just what is going on because the Münster system expresses variation, where possible, as a series of parallel streams of text. Generally, the Münster system tended toward longer variation units (typically of two or more words), especially in cases involving transposition (as here), while ours tended to shorter variation units (typically, one word only). In the Münster system, one can see at a glance that the variation is really based on a set of transpositions. In our system, this takes some effort to discern for the same variants. On the other hand, our system appears to give a finer level of detail about the agreement between the manuscripts than might emerge from the Münster collation. Thus, our format reveals that the text forms b and d here both agree in having the first word as κάθον.

We could see that the Münster system of longer variation units often gave a clearer picture of the different versions of the text (which, of course, is what most readers are interested in). At the same time we thought that for the research we did, seeking to show how the manuscripts are related on the basis of analysis of the agreements they share and do not share, the finer-grained information that we give based on shorter variation units is valuable. Accordingly, in the General Prologue collation published in 2000 we sought a compromise (Solopova 2000). Where it seemed that using a longer variant unit (all three words, in this example) would give a clearer view of the different forms of the text, and would not significantly affect our ability to show agreements at the level of words and short phrases, then we would express variants in those forms of the text in longer phrases. But alongside this, where in some forms of the text the variation seemed better expressed in terms of single words and shorter phrases, then we would use the variation in single words or short phrases.

This procedure led to collations such as the following, for the base text ‘Wel koude he sitte on hors and faire ryde’ in line 94 of the General Prologue (slightly simplified here):

Wel] *39 witnesses*
koude] *39 witnesses*
he] *39 witnesses*
sitte] *39 witnesses*
on] *26 witnesses*; on a *7 witnesses*; a *5 witnesses*; on [an] *1 witness*
hors] *37 witnesses*; horsbak *1 witness*
hors and faire] a goode hors and *1 witness*
and] *33 witnesses*; and therto *5 witnesses*
faire] *36 witnesses*; wel koude he *1 witness*; faire therto *1 witness*
ryde] *39 witnesses*

Here we have the longer phrase 'hors and faire' with a variant 'a goode hors and', that seemed best expressed as a longer variant. However, alongside that phrase variant, and overlapped by it, are a series of variants based on the single words 'and' 'faire.'

This, it seemed to us, gave something of the best of both worlds: longer variants when that seemed clearer, shorter variants when that seemed to give more information.

However, this satisfaction was disturbed by the second event of 1998. In this year we began to collaborate systematically with researchers in the Department of Biochemistry in Cambridge, notably Christopher Howe and Adrian Barbrook, on the application of methods drawn from evolutionary biology to studies of textual traditions. This collaboration was marked by our first joint publication, a paper 'The Phylogeny of the Canterbury Tales' in *Nature* (Barbrook 1998). From the first, this encounter with professional scientists, whose working days are spent considering exactly how evolutionary biology analytic programs work and how data should be prepared for them for the best results, has proved exceptionally fruitful. They paid much attention to finding out exactly how we gathered our data and how we prepared it for analysis. In turn, this caused us to think more carefully than we had about how we do what we do. This process was sharpened in 1999 and 2000 when Barbara Bordalejo (De Montfort, Leicester) and Matthew Spencer (Cambridge) became involved in this work.

Our collaborators on this, and especially Matthew Spencer, pointed out that there was a considerable flaw in our procedure of overlapping variants. The problem is this: manuscript A has the longer reading 'a goode hors and'. Manuscripts B and C have readings on the words 'hors' 'and' 'faire' overlapped by this variant. We can say what readings B and C have at each of these three words. But what reading does A have at each point? And, to put the situation in the reverse: we can say for that for the phrase 'hors and faire' manuscript A has the reading 'a goode hors and'. But what readings do manuscripts B and C have here?

The Cambridge group pointed out that the effect of this problem was that one could not make any statement at all, in cases such as this, about the relationships between A (on the one hand) and B C (on the other). One might find a way about this in this fairly simple case where the single word lemmata ('hors' 'and' 'faire') are all contained within the longer phrase lemma ('hors and faire'). We spent some time considering how this might be done through some process of atomization of the longer variant strings or concatenation of the shorter. But no such solution appeared possible in cases such as that given in the variants on the phrase 'wedded newe a' in the first example. Here, we have one set of variants on the phrase 'wedded newe' and a second on the phrase 'newe a', as well as variants on each individual word. If manuscript A has a variant on 'wedded newe' and B has one on 'newe a' there simply seemed no way one could compare the text of A and B directly, and make any statement at all about the relationship between A and B at those points.

The efficacy of phylogenetic tools is enhanced when the data on which they work contains the most complete possible statement of agreements and disagreements within the population of objects analysed. The effect of our use of overlapping units of variation was that, at every point of overlap, there would be manuscripts for which we would not be able to make any statement about the relationship between them. And, we had very many points of overlap.

At the same time as we were learning of the negative effects of our tolerance for 'overlapping variants' on phylogenetic analyses, we were becoming disenchanted with it for separate, but related, reasons. Our system was excellent at showing, for each variant, just how any one version differs from the base text. Further, where two versions present different readings for the same span of base text, we could also say exactly too how those versions differ from each

other. That is, we could compare directly the reading of manuscripts B and C at 'hors' as both have this word as the lemma for their reading. But where the versions differ from different spans of base text, then our system could not say at all how they differ from each other. Thus, one could not say for any word in the phrase in A 'a goode hors and' just what the reading in B and C is at that word. To put it another way: we would like to be able to point at any word in any manuscript and say: what readings do the other manuscripts have at this point? But this was exactly what our system could not do. With our system, we could only say: at this word, the base text has such and such. We could not always say: at this word, here are all the readings found at this point in all the other texts. Similarly, we wanted to be able to compare any two (or more) manuscripts word by word, showing exactly how they differ. Once more, this system could not do that: we could only show how they severally differed from the base text, not how they differed from each other.

Furthermore, we had grown dissatisfied with the prominence these procedures gave to the base text. This base text was neither the transcript of any one document nor a deliberately-conceived edited text. It was an artificial construct, designed only to permit efficient collation. Yet, our collation method of referring all variants to this base text both severely limited how we could display variation and drew the reader's attention directly to this base text itself. We would have preferred the base text to be quite invisible: instead, it was all too visible.

The cure for removing overlapping variation is simple, but drastic: refer all variants to the same base lemma. That is, the unit of variation has to be fixed by the longest variant present at any point. In the case of the General Prologue example, this will be set by the variant 'a good hors and'. Thus, instead of having four lemmata at this point ('hors' 'and' 'faire' 'hors and faire') with the different manuscripts referred to the different lemmata there will be just one lemma, as follows. Instead of

hors] *37 witnesses*; horsbak *1 witness*
hors and faire] a goode hors and *1 witness*
and] *33 witnesses*; and therto *5 witnesses*
faire] *36 witnesses*; wel koude he *1 witness*; faire therto *1 witness*

We have:

hors and faire] hors and faire *29 witnesses*
- hors and therto faire *5 witnesses*
- a goode hors and *1 witness*
- hors and wel koude he *1 witness*
- hors and faire therto *1 witness*
- horsbak and faire *1 witness*

In the case of the Miller's Tale example, in the sequence of variants on the words 'wedde newe a' there is one set of variants on the sequence 'wedde newe' and another set on the sequence 'newe a'. As these two sequences overlap each other the only solution is to combine the two into one, referring all variants to the single three word sequence 'wedde newe a' rather than the five sequences 'wedde' 'newe' 'a' 'wedde newe' 'newe a'. This is the collation given by our previous system, wirh five different lemmata:

wedded] *53 witnesses*; E wedded *1 witness*
wedded newe] newe wedded *1 witness*, newli wedded *1 witness*
newe] *26 mss*; newly *1 witness*; omitted *1 witness*
newe a] a newe *23 witnesses*
a] *30 witnesses*

Now, this is the collation given by parallel segmentation, with just one lemma:

wedded newe a] wedded newe a *25 witnesses*
- wedded a newe *23 witnesses*
- newe wedded a *1 witness*
- E wedded newe a *1 witness*
- wedded newly a *2 witnesses*
- newli wedded a *1 witness*
- wedded a *1 witness*

3. Parallel segmentation collation

We call this 'parallel segmentation' collation, because the collation of the whole text is composed of a sequence of segments, one after another, and in each segment all various readings are presented in parallel with each other. There are two points to be observed here. Firstly, it is possible to compare the reading of any one manuscript within any sequence with the reading of any other manuscript within the same sequence. This removes the problem of missing information in phylogenetic analysis and also makes possible displays which, for example, show the reader who is checking the reading in any one manuscript just what are the range of readings corresponding to that in any other manuscript. Thus, the boxes which appear over each word in every witness transcript on this CD-ROM showing the variants at that point in all witnesses, are built on this parallel segmentation collation. Also built on this collation are the 'compare' views, which allow the reader to compare any two witnesses, and the 'select witnesses' feature in the word-by-word collation view, which allows the reader to include or exclude any witnesses from the collation.

The second point is that in this parallel segmentation format, once the collation is done we can set aside the base text without any loss of information at all. Thus, each collation might be presented simply as:

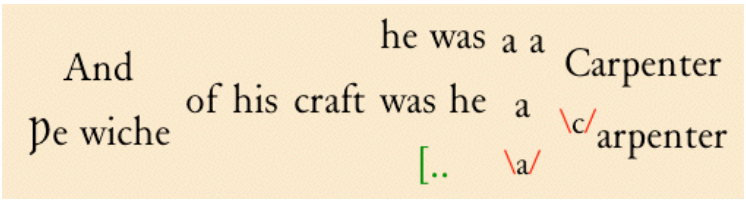
hors and faire *29 witnesses*
hors and therto faire *5 witnesses*
a goode hors and *1 witness*
hors and wel koude he *1 witness*
hors and faire therto *1 witness*
horsbak and faire *1 witness*

and

wedded newe a *25 witnesses*
wedded a newe *23 witnesses*
newe wedded a *1 witness*
E wedded newe a *1 witness*
wedded newly a *2 witnesses*
newli wedded a *1 witness*
wedded a *1 witness*

Therefore, there is no need to give the base text at all. Both the Wife of Bath's Prologue and the General Prologue CD-ROMs opened with the 'Base text for collation', and all collation views employed the base text as the reference for all comparisons. This CD-ROM opens with our transcript of the Corpus, Oxford manuscript and this is used as the default reference for all collations. We chose this manuscript partly because of its likely early date, partly because of the excellence of the images available to us, but very much because we wish to remind readers that there are many other manuscripts of the *Tales* beside famous pair Ellesmere and Hengwrt. The reader may choose to open with any other witness, and to use any witness as the reference for all collations.

The use of this parallel segmentation method has other benefits. Manifestly, it is much easier to comprehend the different forms of the text when all the variants are presented in parallel in this manner. One can also create views such as this, for line 3 of The Miller's Tale:



Here, the collation is presented as a series of segments, corresponding to the standard text 'And' 'of' 'his' 'craft' 'he was' 'a' 'Carpenter'. Where there are variants they are stacked one above the other. A further benefit is that we are able to make this one collation serve all our needs. For the General Prologue CD-ROM, we had one collation for the regularized collation, another for the unregularized collation, a third for creating the phylogenetic analysis, a fourth for making the variant database. We do all these things, in this CD-ROM, from a single collation.

It has taken some time to understand all the implications of this new collation format and to implement it. Core segments of the collation software we use, the *Collate* program, had to be rewritten and tested. We were fortunate indeed in our partners in two other major text edition projects, both of which now employ the same parallel segmentation method: the edition of the Greek New Testament in Münster, and the edition of Dante's *Commedia* led by Prue Shaw of University College London. People in both these projects, and especially Klaus Wachtel from Münster, spent many hours studying collation output with me, and working out how it should appear, and what would be needed to make a collation which both represented the textual differences in the most lucid possible form, but would also support all the kinds of display and data analysis which we wanted.

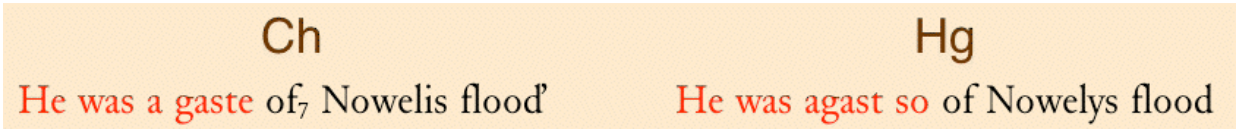
4. The next step

Although parallel segmentation collation is unquestionably a considerable advance on our previous collations, it is not the end of the matter. Its effectiveness is related directly to the length of the variant segments: the shorter the variant segments, the better it is at showing the precise differences between manuscripts; the longer they are, the more likely it is that correspondences between the manuscripts within the longer variant strings will be obscured. When the segments are just one word or two words long, as in the instance from line 3 of The Miller's Tale given above, the results are quite satisfactory. But consider the variants on the first four words of line 646 of The Miller's Tale, 'He was agast so of Nowelys flood'. The seven witnesses of the important b group all transpose the fourth word 'so' to the beginning of the line. Accordingly, we are constrained by parallel segmentation to make the segment four words long, and refer all the witnesses to this longer segment. Thus, we see:

He was agast so *33 witnesses*
He was agast *4 witnesses*
So he was agast *6 witnesses* (all b group)
He was so agast *7 witnesses*
He was agast and feerd *2 witnesses*
So was he agast *1 witness* (b group)

Just as a presentation of the variation at this point, this is quite efficient. But as a representation of the exact linkages between the witnesses, it is rather inefficient. These six variants are presented in simple parallel, as if no two of them are any closer than any other. But manifestly, that is not true. The second and fourth readings 'He was agast' and 'He was so agast' are much closer to the first reading 'He was agast so' than they are to either the third and sixth readings. In turn, the third and sixth readings 'So he was agast' and 'So was he agast' are much nearer each other than they are to the other readings. One can see this very clearly on the variant map for this point, where the single witness Wy with 'So was he agast' (coloured brown) is shown as a b witness and clusters with the other six b witnesses (coloured green), which all have the reading 'So he was agast'. Similarly, one can see from the variant map that the 44 witnesses which contain the first, second and fourth readings (coloured red, blue and purple) all cluster together.

Indeed, information about these links between the first, second and fourth readings on the one hand, and the second and sixth on the other hand, would have been very useful in aiding the phylogenetic programs to build the picture of witness relationships shown in the variant map (see next section, for description of how we make this map). But in our current implementation of the parallel segmentation method, we do not attempt to identify any such correspondences between the different readings in any segment, and so this information is not available to the phylogenetic programs. Similarly, this referencing of what may be a difference in just one word to a longer parallel segment can lead to counter-intuitive displays when we are comparing just two manuscripts. Here is the comparison of the Christ Church manuscript (one of the four with the second reading) with the Hengwrt manuscript (the first reading) for this line:



In fact, the manuscripts differ substantively only in the presence of 'so' in Hg and its absence in Ch. One would expect then that only this difference would be shown. But instead, the whole phrase 'He was a gaste' in Ch is shown as a variant

on the whole phrase ‘He was agast so’ in Hg. This is a direct result of the parallel segmentation method as we currently employ it. Once it has found the segments, the collation stops and just presents the segments it has found. In this collation system, all variants at any point are equally unlike.

One can imagine a collation system which does not stop at the point where it has identified the parallel segments, but actually carries on within the segments, seeking to link them at a finer level of detail. This set of variants might be expressed in three sequences. First, this sequence for the first, second, fourth and fifth readings:

He was		[46 wits.]	*the first, second, fourth and fifth readings*
	so agast	[7 wits.]	*the fourth reading*
	so agast and feerd	[2 wits.]	*the fifth reading*
	so agast	[37 wits.]	*the first and second readings*
	so	[33 wits.]	*the first reading*
	-	[4 wits.]	*the second reading, flowing from the first*

Second, this sequence for the third and sixth readings:

So agast	[7 wits.]	*the third and sixth readings*
he was	[6 wits.]	*the third reading*
was he	[1 wit.]	*the sixth reading*

Finally, we need a third sequence to link together the first and third readings, as they differ only in their placement of the word ‘so’. Effectively, this sequence then acts to link together the whole first sequence (containing the first reading) and the second sequence (containing the third reading), so joining all fifty-three witnesses active at this point.

He was agast so	[33 wits.]	*the first reading*
So he was agast	[6 wits.]	*the third reading*

One can see, conceptually, both how this would give much more, and much better, information to phylogenetic tools, and how this would lead to much more precise displays of the exact differences between the manuscripts. For example: one could extract from this the information that Ch agrees substantively with Hg in the three words ‘He was agast’ but differs from it only in the absence of the fourth word ‘so.’

Making apparatus on this more sophisticated model presents significant challenges. Again, one could conceive a software tool which could decompose the six variant sequences for line 646 using parallel segmentation into the three variant sequences, two of them with further sub-variation. But writing such a tool will not be easy, especially as it will be necessary to allow (as with the current *Collate* tools) the scholar to over-ride what the program finds. Presenting these apparatus will also be rather difficult. It has taken us a great deal of experimentation to arrive at the presentation of parallel segmentation apparatus offered in this CD-ROM, and this projected system will provide a more complex apparatus than that.

By definition, this problem only affects variants on segments of more than one word. As part of the conversion process by which the *Collate* apparatus is converted into the nexus file format used for phylogenetic analysis we are given full lists of all types of variants. From these lists, we can deduce that in the whole apparatus for Link 1 and The Miller’s Tale we have variants as follows:

One word variant segments: 5106
Two word variant segments: 355
Three word variant segments: 85
Four word variant segments: 26
Five word variant segments: 10
Six word variant segments: 4
Total variants on words: 5586

That is: of the mass of variants in the Miller’s Tale, less than 10% (some 480 of a total of 5586) are variants which might require this further treatment. As these variants are indeed greatly outnumbered by the single-word variants, one could

argue that we are gilding the lily: that the information already available is quite sufficient for valid results. However, it is certainly true that the picture of witness relations we derived from the phylogenetic analysis based on the parallel segmentation apparatus is much clearer than anything we had for the General Prologue and the Wife of Bath's Prologue. Undoubtedly, this advance is largely (if not entirely) due to the adoption of parallel segmentation. Improving even further the data we provide, by the refinement here suggested, may lead to still clearer views of the relations among the manuscripts. Further, the relatively smaller number of such variants suggests that some kind of manual tool might be possible, whereby the editor could efficiently determine at these points how the variation should be expressed.

We will continue to examine this matter. Later publications from the project may seek to implement this more refined version of the parallel segmentation algorithm.

5. Notes

1. See Whittaker 1991 for a discusson of collation. However, Whittaker concentrates on issues relating to physical methods of hand collation (the formats of paper to use, how to record variants) and says nothing about exactly what should be recorded in a collation, I am conscious of a debt to Michael Sperberg McQueen for many of the ideas relating to collation formats raised at the end of this article: cf. his discussion (with Claus Huitfeldt) in http://helmer.aksis.uib.no/claus/mlcd/papers/texmecs.html.